DOCUMENT RESUME

ED 180 766 SE 029 384

TITLE	Proceedings: Summer Conference for College Teachers on Applied Mathematics, University of Missouri-Rolla, 1971.
INSTITUTION	Committee on the Undergraduate Program in Mathematics, Berkeley, Calif.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	71
GRANT	NSF-CUPM-G-23827
NOTE	390p.; Contains light and broken type
EDRS PRICE	MF01/PC16 Plus Postage.
DESCRIPTORS	Algebra; *Calculus; *College Mathematics; *College Teachers; *Conference Reports; Higher Education; *Mathematical Applications; *Probability; Statistics; Teacher Education
IDENTIFIERS	Differential Equations; *Linear Algebra

ABSTRACT

Proceedings from four sessions of the Summer Conference for College Teachers on Applied Mathematics are presented. The four sessions were: (1) Applications of Elementary Calculus, (2) Applications of Linear Algebra, (3) Applications of Elementary Differential Equations, and (4) Applications of Probability and Statistics. Nine lectures were given in Session I including "Where Calculus Teaching Fails," and "Miscellaneous Applications and Continuations." Sessions II and III included 14 lectures each. Among those in Session II were "An Electrical Ladder Network," and "Markov Chains." Session III included "Applications of Differential Equations in Biology," and "One-Dimensional Eigenvalue Problems." Session IV contained 17 lectures including "Binomial Probabilities and Applications," and "Elementary Bayesian Statistics." (MK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED180766

REC'D [illegible] JAN 8 1973

To Archives
File: CUPM
G-23827

PROCEEDINGS

SUMMER CONFERENCE
FOR COLLEGE TEACHERS
ON
APPLIED MATHEMATICS

University of Missouri – Rolla
1971

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Mary L. Charles
NSF

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

SE 029 384

Supported by
NATIONAL SCIENCE FOUNDATION

January 1973

Additional copies of these proceedings may be obtained from CUPM,
Post Office Box 1024, Berkeley, California 94701.

PREFACE

The Summer Conference for College Teachers on Applied Mathematics was held at the University of Missouri--Rolla during the summer of 1971. This conference was planned with the active cooperation of the Mathematical Association of America's Committee on the Undergraduate Program in Mathematics (CUPM) and the CUPM Panel on Applied Mathematics. CUPM feels strongly that material of the type contained in these proceedings be given wide distribution, for this material should contribute significantly toward improving the effectiveness of undergraduate mathematics teachers. Both CUPM and the Panel on Applied Mathematics applaud this effort to acquaint college teachers with applications of mathematics and encourage other institutions to organize similar programs.

Conference Director

Glen Haddock

Senior Lecturers

Charles Antle
Thomas F. Banchoff
Wilfred Kaplan
Ben Noble

Visiting Lecturers

Leroy Folks
Robert Ritt
J. Laurie Snell

Recorders

Charles Hatfield
Merrill Hume
Otho R. Plummer, Jr.
Lyle E. Pursell

Edited and prepared for publication
by the
CUPM Central Office staff

Andrew Sterrett, Executive Director
Paul J. Knopp, Associate Director
William E. Mastrocola, Associate Director
Katherine B. Magann, Administrative Assistant

TABLE OF CONTENTS

Session I. APPLICATIONS OF ELEMENTARY CALCULUS

Senior lecturer	Thomas F. Banchoff
Recorder	Charles Hatfield

Lectures by Professor Banchoff

1. Introducing Calculus (Part 1) 1
2. Introducing Calculus (Part 2) 7
3. On Extracting Simple Problems from the More Complex 13
4. Where Calculus Teaching Fails 21
5. Miscellaneous Applications and Continuations 27
6. An Arch, A Cup, and e 32
7. The Playpark Problem 38

Lectures by participants

8. Newton's Law of Gravitation [Bernard Rasof] 44
9. Comments on the Teaching of Calculus and its Applications to Students in the Life and Management Sciences [Clifford W. Sloyer] 54

Session II. APPLICATIONS OF LINEAR ALGEBRA

Senior lecturer	Ben Noble
Visiting lecturer	J. Laurie Snell
Recorder	Otho R. Plummer, Jr.

Lectures by Professor Noble

1. Introduction 60
2. An Electrical Ladder Network 62
3. Numerical Solutions of Linear Equations 68
4. Generalization from One Variable to n Variables; Norm 75
5. Applications of Systems of Differential Equations with Constant Coefficients 83
6. Diagonalization and Quadratic Forms 93
7. Linear Programming 102
8. Application of Matrices to Economic Models and Social Science Relationships 111
9. Linear Least Squares Method; Regression 120

Lecture by Professor Snell

10. Markov Chains 134

Lectures by participants

11. A Mechanical Vibrating System [Bernard Rasof] 150
12. Algebraic Coding Theory [Alan E. Rosenberg] 178
13. Relativistic Quantum Mechanics [Otho R. Plummer, Jr.] 186
14. Artificial Inbreeding [Ronald A. Knight] 188
Bibliography 193

Session III. APPLICATIONS OF ELEMENTARY DIFFERENTIAL EQUATIONS

Senior lecturer	Wilfred Kaplan
Visiting lecturer	Robert Ritt
Recorder	Lyle E. Pursell

Lectures by Professor Kaplan

1. General Lecture 199
2. Feedback Control Systems 208
3. Air Resistance, Higher-Order Systems 216
4. Existence Theory 223
5. Electrical Circuits and Networks 231
6. Differential Equations in Economics 238
7. Applications of Differential Equations in Biology 245
8. Remarks on Centers, Boundary Value Problems 252

Lectures by Professor Ritt

9. Simple Pendulum and Two-Particle Problems 258
10. Predator-Prey Model, Van der Pol Equation, Poincaré-Bendixson Theorem 268
11. One-dimensional Eigenvalue Problems 271

Lectures by participants

12. Mass Action [Richard R. Tebbs] 276
13. Heat Transfer [Richard R. Tebbs] 279
References 283

Session IV. APPLICATIONS OF PROBABILITY AND STATISTICS

Senior lecturer	Charles Antle
Visiting lecturer	Leroy Folks
Recorder	Merril Hume

Lectures by Professor Antle

1. Binomial Probabilities and Applications 286
2. The Poisson Probability Law and Applications 291
3. Bayes' Theorem and Applications 293
4. Models and Estimation 302
5. Confidence Intervals 307
6. Contingency Tables 311
7. Sampling Problems 316
8. Location and Scale Parameter Families 320
9. An Application of Markov Chain Theory 326
10. Elementary Bayesian Statistics 330
11. Regression Analysis, an Example 336

Lectures by Professor Folks

12. Analysis of Variance 339
13. Some Classes of Problems in Statistical Methods 345
14. Regression Analysis 353
15. A Theory of Statistical Inference 361

Lectures by participants

16. Modeling of Tree Nut Removal [Jaw-Kai Wang] 367
17. A Telephone Exchange Problem [Andreas Thuswaldner] 377

CALCULUS

Lecture 1: Introducing Calculus (Part 1)

As teachers we form a bridge between calculus and our students--more modestly, one of the bridges, for undoubtedly students also learn from textbooks and from fellow students. This fact of our intermediate position suggests two questions: What situations lead to applications of the calculus? What are our students interested in? The answers to these questions should suggest the areas of our bridge-building operation. Realistically we must start from where they are and then take them as best we can to where we want to go.

To answer the second question, I find a brief questionnaire helpful. I write my name on the board and tell the students something about myself: where I was born, where I went to school, where I live at present, etc. Then I request from them on paper similar information, which I then memorize in order to have a composite picture of the class, but a picture that is composed of specific and detailed information. (The questions are: Name? Special interest? [Major and minor, or equivalent ideas.] Where are you from? What mathematics courses have you taken? What other courses have you taken? Why are you taking this course? and/or What do you expect to gain from it? All of this helps them to get involved immediately with calculus and to express what they expect from it.)

I tell them why I am giving the course and what I expect to do in it. I tell them what I hope the students will derive from it.

The makeup of the class helps orient me to their needs. A class consisting of arts and science majors will likely have a different set of interests from one consisting of mechanical and civil engineers. Collection of this data by the teacher is also reassuring to the student; he knows that his teacher cares about his name and his goals as well as his university number and that he can expect further to be treated like a person. It suggests that the teacher is humane in making an honest effort to begin roughly where the student is.

I also tell the students where I get my problems. Some of the problems come from newspapers. For example, I have here a copy of the June 3 local newspaper. Here is the way I use it to introduce some key ideas of the calculus.

Temperatures on June 2

T(2) = 50
T(8) = 59
T(14) = 71
T(20) = 61

Then I ask a series of questions and get student responses somewhat as follows:

Was it ever 40?	I don't know.
Was it ever 55?	Probably.
Was it ever 50?	At least once.
Was it ever 65?	At least twice.

I question each response to see what argument they have in mind to support their answer (even though I might think their response is all right). All this is assumed to take place at the level of a normal (inquisitive, but not too knowledgeable) freshman who is willing to play a little game with you, as teacher.

Each of these questions will help the students to raise the questions of what the data mean, how they were obtained, and how they got into the present form (round-off or chop-off). The airport could be called upon for further information along this line. I generally discourage the word "continuous" and its cognates. If the student should use the word, I ask him for an intuitive translation of it. The name of the game is "intuition and geometry" at this point.

To help answer the first question, further data is now supplied: the high and low for the date in question.

High at t = 15: 71; Low at t = $4\frac{1}{2}$: 47

Now one has the range of the function (as far as the tables go) and he can answer questions like "Was it ever 40?" or "Was it ever 100?"

Phase Two. I now provide data on the humidity for the same date:

Humidity on June 2

H(2) = 80
H(8) = 65
H(14) = 46
H(20) = 53

Now we can ask a similar set of questions about humidity, using, say, 55, 40, 90, and 105. (If someone calls attention to the last number, fine.) Further questioning occurs about the units in the cases of T and H. This also provides an opportunity to draw out the "operational definition" which one uses in this sort of determination. (Reference is here made to Bridgman's now-classic formulation of the definition of physical concepts such as temperature and humidity.)

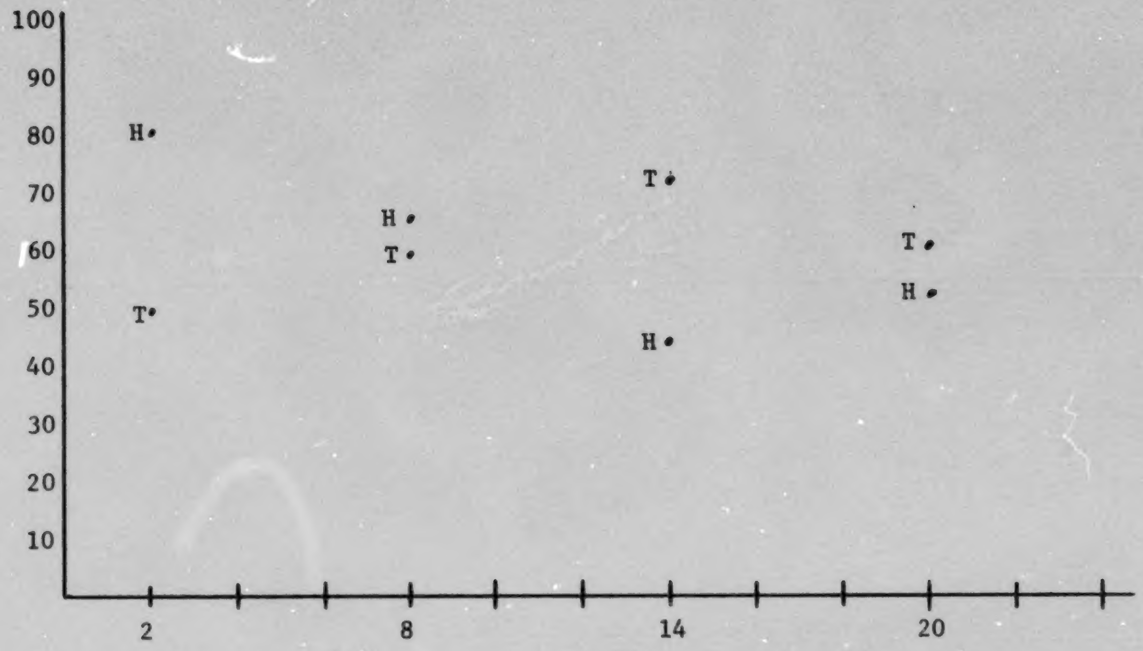
We can also discuss the domain of each of the functions, $0 \leqq t \leqq 24$, and the ranges. Attention is called to the number 105 for a value of H, and questions are designed to bring out the impossibility of that datum.

Phase Three. We turn to questions involving both T and H. One "natural" question is whether T and H were ever the same: Is there a value of t_o in the domain $0 \leqq t \leqq 24$ for which $T(t_o) = H(t_o)$? After some discussion to assure the students that it is the numerical values of T and H which we have in mind (disregarding units), it is a good idea to take a vote on the question. Notice that asking for a head-count creates a little more pressure on the student to get involved in the problem, in case there is any reluctance to think about it. You can expect a division of the house on the question.

If there is some objection as to whether the question is a natural one, one can ask them about physical situations in which two functions do have the same value for some significant event associated with them. One example many students will suggest is that of air locks (re space travel in science fiction). Another is the locks in a canal. (For example, the Panama Canal's three locks on either side of Gatun Lake.) A third example is the problem of when to escape from a car that has plunged into water deeper than the car.

A follow-up on representative votes on either side is a good idea, for it allows the instructor to determine the quality of back-up reasons for the votes. Consider one vote from either camp, and draw out the voter's reason for his opinion. (What led you to a no-answer? What led you to a yes-answer?) Here one can talk about the notion of proof as simply a convincing agrument--all in the context of the calculus class. Other subject matters might well arise from other contexts, and the rules of the game will vary somewhat. But "proof" is still something convincing in the setting of the agreed subject matter. The students will probably agree that it seems a reasonable ground rule for the course.

Phase Four. Ask for other ways of presenting the data. Someone will usually suggest the idea of a two-dimensional representation in a graph. For the above data, for example, consider the following graph(s):



Now we again ask the questions ("Was it ever 40?", etc.) and this time we use the graphs to make plausible answers and show the variation allowed by and consistent with the given data.

There are other methods of presenting the data. One student thought of the question whether T and H were ever the same in terms of a horse race. At one time the humidity-horse is ahead and at another time the temperature-horse is out front. So the T-nag must have somewhere passed the H-nag. Here is a solution which is implied by the given data. This problem also provides an excellent opportunity to distinguish between the existence of a solution and the solution itself. (The former is what the data imply.)

If we wish to narrow further the margins on the question "When do the graphs cross?", we might call the airport and obtain the crossing to within one hour.

Phase Five. A suggested generalization of the problem. Since there were other phenomena cited (air locks, water locks, underwater escape from car), there seems to be a principle that is wider than mere (artificial?) comparison of temperature and humidity. How can we formulate the problem more generally? One can deal with primitive terms if he likes, but it helps intuitively to have the notion of intermediate-value property. Give examples of it. (Perhaps some of the students have used the idea already in their high school algebra course if it included a technique for finding real irrational roots of polynomial equations to arbitrary accuracy. There the principle is usually assumed without any effort at proof.) From these empirical roots one can then draw out a definition:

Definition. A function F with domain $a \leqq x \leqq b$ has the intermediate-value property (IVP) if, whenever C is between F(a) and F(b), there is at least one x_o in the domain of F such that $F(x_o) = C$.

Now we return to our problem. Our assumption about the functions T and H is that each has the IVP. We can restate our conjecture about T and H:

Conjecture. If F, G have the IVP over $a \leqq x \leqq b$ and if $F(a) < G(a)$, $F(b) > G(b)$, then there exists a value x_o in the domain such that $F(x_o) = G(x_o)$.

This conjecture is then left to the students to prove. Some will give up on it; others will write nonsense; still others will think it obvious and will be a little mystified about what is requested by a "proof" of something that seems so clear. The few (?) who get it should feel a real sense of accomplishment in the success.

Ask the students for further examples of pairs of IVP functions. One can suggest whole categories of functions. Without too much coaxing they should come up with constant functions, then linear functions, then perhaps quadratic functions. Then you can suggest conjectures such as: If $y = x$ is IVP, then so is $y = mx$; or If $y = f(x)$ is IVP, then so is $y = f(x) + c$; or If $y = f(x)$ is IVP, then so is $y = cf(x)$; or If $f(x)$ and $g(x)$ are both IVP, then so is their sum $f(x) + g(x)$.

Once the idea is clear about the problem that has been generalized, one can give further pairs of functions, the object being to solve for the "magic x" that makes the two functions have a common value; this can usually be done even though the student does not find it within his power to prove the conjecture.

To generalize this notion, it is good to distinguish for the students the problems they are expected to do from those they are invited to think about.

Further examples of pairs of functions to be "coincidentalized":

1.	$f(x) = 2x + 3$	$g(x) = 5 - x$	$-4 \leqq x \leqq 4$
2.	$f(x) = x^2 + 3x + 1$	$g(x) = 5 - x^2$	$-4 \leqq x \leqq 4$
3.	$f(x) = x^3 - 3x$	$g(x) = 2$	$-4 \leqq x \leqq 4$
4.	$f(x) = x^3 - 3x$	$g(x) = 18$	$-4 \leqq x \leqq 4$
5.	$f(x) = x^2$	$g(x) = 2^x$	$-4 \leqq x \leqq 4$

If someone asks about the meaning of 2^x, suggest that he think about it and try to come up with a reasonable meaning, and use it in the problem. In any case, assure him that the function will not be forgotten, but will be discussed further.

Lecture 2: Introducing Calculus (Part 2)

Consider some more data from a newspaper:

$F(8) = 46.25$	High 46.625
$F(17) = 46.50$	Low 45.75

Some questions are then asked, and (educated) guesses are invited; for example, "Was it ever 46.27?", "What are your assumptions?" Some think "it" refers to temperature or humidity. The time 17 (5 p.m.) suggests perhaps something other than temperature or humidity, unless the datum was suppressed in the previous lecture. Further discussion of assumptions is encouraged if the group seems to need it. A young investor in the class might be able to guess the source of the data. Actually, it's the stock market. The form of the data as they are commonly given (in eighths) would likely give it away, so the data are given here in decimal form. Actually, we would expect that the form of the data might give us some information about the kind of data.

Once the stock market is mentioned, some will know immediately that 46.27 is an impossible quotation. Those not familiar with the stock market might feel a little left out of this example. The given data are graphed. It is clear that we now have on our hands a step function. The range is only a finite number of values which includes the high and low for the day (these might even be the same). The domain consists of the relevant hours of market business, a fraction of the 24 hours of the day. Axes should be drawn so that the instructor can discuss with the students the various possibilities for graphs of step functions compatible with the data.

It is pedagogically desirable to discuss at this juncture a discontinuous function such as that of the stock market, lest it be inferred from the previous lecture that all functions are continuous or all functions have the IVP. After the discontinuous function above is discussed for some time, it is a good idea to consider a discontinuous function which is not a step function. Again we must be constantly on guard lest the student draw false conclusions from our examples. After all, we give him examples precisely so that he can draw proper deductions from them. But his mind and attitude are not on the same track as ours. We want not only to be understood but also not to be misunderstood.

Students should be encouraged to come up with their own examples of functions continuous and functions discontinuous. In this particular context it is the examples of the latter that are especially welcome. One of my favorites is that of "putting and putting." One "putting" refers to golf balls and the other to the shot. Presumably each involves a continuous effort (which we might call inputt and input!), but the result (outputt and output, respectively) is different. For the case of golf, the putt is a failure unless the ball goes into the cup. So scoring is 0 or 1, in a sense. The distance of the shot put, however, could be a continuous range.

Other examples: (1) Raindrops falling on the continental divide. Here one must make a definition of raindrop somewhat different from the ordinary one, to insure the proper dichotomy in destination of the drop. (2) Acceleration on a piece of chalk that rolls on a horizontal table to the edge and then falls off. Acceleration is zero until the chalk falls off, at which point it is the gravitational constant (with proper sign) from then on. We assume here that there is no friction on the table top serving to decelerate the rolling chalk. (3) Thermostatic control, or other servo-mechanism, offers a similar mechanism with a discrete range. Some cars have a system that activates the air conditioner or the heater when the threshhold temperatures are reached. This gives a three-state range.

Some further examples follow:

A. A sprinter gets mixed up with a gambling ring. They kidnap him, make threats, but give him a "fighting chance." He stands in front of a car that is about to race him. There is a pit or pool at the 60-yard mark. If he reaches it first, he is safe. If not, he is thereby "rubbed out." The car is known to be able to go from 0 to 60 m.p.h. in 10 seconds flat. Does the sprinter make it? Here the student is encouraged to make reasonable assumptions but is not told what to assume other than the given problem statement. This is typical of a class of problems that demand some physical assumption(s) in the process of solution. Regarding problems of this type, Agnew states, in connection with his now-classic snowplow problem, "Much of the progress in science is due to men who have the courage to make assumptions, the good sense to make reasonable assumptions, and the ability to draw correct conclusions from the assumptions." [Agnew, Ralph P. Differential Equations, 2nd ed. New York, McGraw-Hill Book Company, 1960. p. 39]

B. If you travel 120 miles in 2 hours, then at some time you were traveling at 60 m.p.h.

C. A boy holds a ball in his hand at two different times. You did not see what he did with it, if anything, in the interval between. Conclusion: At some time it must have had zero velocity. (Similar problems were suggested for trains and for airplanes.) The intuitive reasoning here is to be elicited from the students.

This kind of example is good, not only to instruct the student in the ideas of calculus which it illustrates but also to assure him in so doing that there is a close connection between geometry and kinematics.

D. Mountain climbing. You climb to the top of one mountain peak in the twelve hours between 6 a.m. and 6 p.m. The next day you return from the peak, traveling at the same hours. Show that you were at the same point on the mountainside at the same time of day in each of your two treks.

E. Crumpled paper is dropped at random on a sheet of paper of the same size and shape. Then at least one point of the crumpled paper is directly above the corresponding point on the uncrumpled sheet.

F. Show that your height in inches and your weight in pounds were identical at one time in your life.

G. Two sailing boats take different paths from A to B, but each path has a continuously turning tangent in the voyage. Are they ever sailing parallel? That is, are the tangents ever parallel at the same time?

On Cake-cutting and Geometry. Assume you have an ordinary round, flat cake, without icing. In the spirit of the foregoing (today and yesterday), we ask the student to cut the cake in half. If someone says he can, ask him how. Now what if the cake is of irregular shape? Can you still halve it? How? What idea do you use? Can you halve it, cutting in a given direction?

Suppose, to make the matter definite, that the cake is a scalene triangle. Can you then halve it? How? To what geometry do you appeal?

Again suppose the cake is square. Can you halve it? (No trouble on this one!) Can you halve it, but cutting in a given direction? (Key: yes, because of the central symmetry.)

Return to the more general problem of halving the irregular cake. Let L(x) represent the area of the cake to the left of the line through x and parallel to the y-axis. The domain of L is the set of all real x. The range of L is from

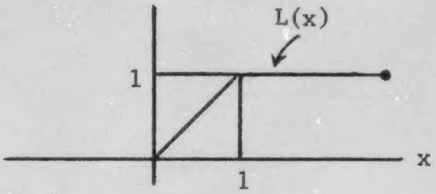
0 to C, the area (proportional to volume) of the cake. So if $y = L(x)$, then $0 \leqq y \leqq C$.

Question: Is there a value of x_o in the domain of L such that $L(x_o) = \frac{1}{2}C$?

Discuss the example long enough for someone to come up with the suggestion that such an x_o does exist because L has the IVP.

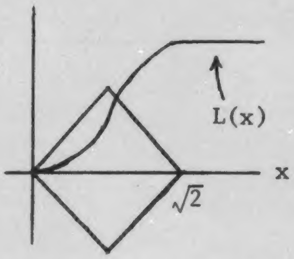
If the bell is about to ring, assign the sketch of L as homework. Discuss several shapes of cake and the corresponding L, drawing a graph for each.

A. Square Cake on Edge



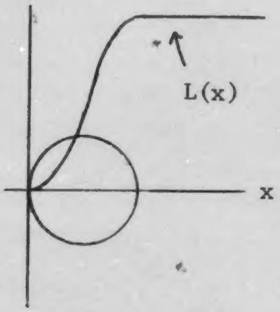
$$\begin{aligned} L(x) &= 0, & x &< 0 \\ &= x, & 0 &\leqq x < 1 \\ &= 1, & 1 &< x \end{aligned}$$

B. Square Cake on Diagonal



$$\begin{aligned} L(x) &= 0, & x &< 0 \\ &= x^2, & 0 &\leqq x < \frac{\sqrt{2}}{2} \\ &= 1 - (x - \sqrt{2})^2, & \frac{\sqrt{2}}{2} &\leqq x < \sqrt{2} \\ &= 1, & \sqrt{2} &\leqq x \end{aligned}$$

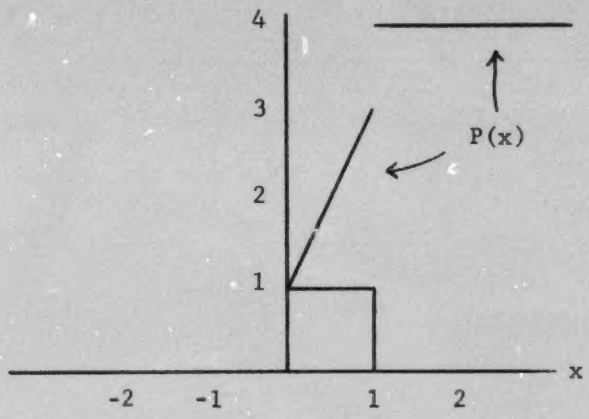
C. Circular Cake



Assign calculation of L(x) for homework.

Now we can look at similar functions for the perimeter. Here we put the icing on the cake! We consider only the icing around the perimeter, the normally iced cake being a modification of this one. Define a perimeter function

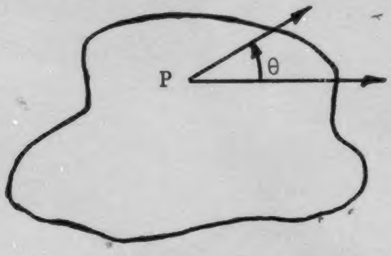
P(x) as follows: let P(x) be the amount of the perimeter to the left of or including x, where x is any real number. Below is the example of a square cake on its edge, with P(x) sketched on the same axes.



$$\begin{aligned} P(x) &= 0, && x < 0 \\ &= 2x + 1, && 0 \leqq x < 1 \\ &= 4, && 1 \leqq x \end{aligned}$$

Problem. Given a cake of irregular shape and a point P on the cake, is there a halving cut that goes through P? (This could be assigned as a homework problem.)

Here we define $L(\theta)$ to be the amount of cake lying to the left of the half-ray making angle θ with an arbitrary line drawn as polar axis, using P as the pole. The domain of L is $0 \leqq \theta \leqq 2\pi$ and the range of L is contained in $0 \leqq y \leqq C$, where $y = L(\theta)$ and C is again the area of the irregular cake. Observe that $L(\theta) + L(\theta + \pi) = C$.

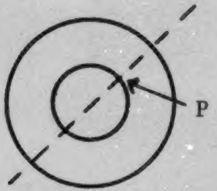


We should bear in mind that while for "reasonable" sets bounded by simple closed curves the center of gravity (c/g) always exists (refer to the definition as the quotient of two integrals, each of which exists under the circumstances we have in mind), the center of area (c/a), a point for which any line through it halves the given set, need not exist. And if c/a does exist it is not necessarily unique. For example, if the set of points consists of the interval from 1 to 2 and its symmetric counterpart with respect to the origin, then any point between -1 and 1 on the line serves as c/a.

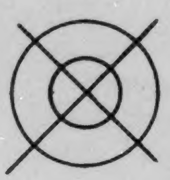
Conjecture 1. If c/a exists, then c/a = c/g.

Conjecture 2. If a figure F has a center of area, then F must be centrally symmetric with respect to c/a.

Further questions: Is there always a point P in a plane figure such that there exist two halving lines through P? (Consider the angel-food cake type--the annular region between two circles of different positive radii. In this case,



each point in the cake has a unique halving line. The center of the annulus (not in the cake) has an infinite number of halving lines passing through it.)



CALCULUS

Lecture 3: On Extracting Simple Problems from the More Complex

We discuss first a precalculus problem to be solved in class after the students have had an opportunity to consider the problem overnight or over the weekend. Again, it has the ring of the common or natural or familiar. It is purposely posed in an underdefined form as follows:

Problem. What is the largest package (rectangular box) that can be sent through the U. S. mail?

If students object to the underdefinition of the problem, it is an opportunity to emphasize the importance of obtaining your own information. Sooner or later they learn that the boundary condition (no pun intended) is that length plus girth must not exceed 72 inches. (The number 72 is not universal, but applies to a good deal of the U. S. Some thought that it could be 100 in some rural localities.) If you don't use the full 72 inches, you can get a larger package by multiplying by an appropriate factor and bringing it up to 72 exactly. So we use length plus girth equals 72 inches. Our problem then amounts to finding the maximum value of $V = \ell wh$, where $\ell + 2w + 2h = 72$.

"We don't know how to start," some will complain. Your reply should be, "That's okay. Go home, generate some data, think about it, try to come up with a solution. A quitter never wins.", etc.

Some will return with a few numbers, others with elaborate tables and some guesses about how big the maximum size package should be. Some will discover from their tables that w and h enter symmetrically into the problem; this suggests that the girth is that of a square cross section. If they see that w and h have the same role in the problem, they will probably agree to substituting one for the other and thus cutting down the number of unknowns. If we fix the length at ℓ_o, then $\ell_o + 2w + 2h = \ell_o + 4h = 72$.

But if they are unwilling to accept the equality of w and h, we can proceed in a different direction with the fixed length ℓ_o. The above equation implies that $w + h = \frac{1}{2}(72 - \ell_o)$, so we maximize the function $V = \ell_o wh$ by maximizing V/ℓ_o. This two-dimensional problem will be familiar to some of the students already, from analytic geometry or even algebra. The constraint in this problem is the linear relation on w and h given above, which leads to

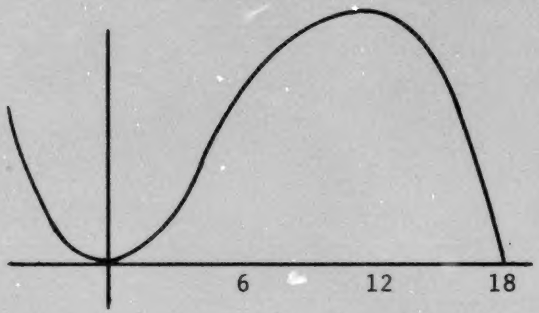
21

$w + h = k\ [= \frac{1}{2}(72 - \ell_o)]$. Immediately we see that V/ℓ_o attains a maximum value at $w = h = \frac{1}{2}k$, since we are essentially dealing with the parabola $y = kh - h^2$, where $y = V/\ell_o$.

Return to the three-dimensional problem. If we accept the equation $w = h$, then the boundary constraint is $\ell + 4w = 72$, so that

$$V(w) = \ell w^2 = (72 - 4w)w^2 = 4(18 - w)w^2. \qquad (1)$$

Again some students are tempted to "throw in the towel" at this point. Some many even try the same trick used in the case of two dimensions, because there is some reinforcement, given the shape of the curve that results from (1). One can defer solution at this point until there is more algebraic aptitude, but if the students are ready algebraically, one can continue in the same precalculus mode. Insistence on finding the solution without using calculus upsets some students, especially those who have already been exposed to the subject. These students will appreciate reassurance at this point that there is some virtue in tying the calculus hand behind one's back.



Since the length cannot be less than either of the other two dimensions, we have some leverage in obtaining another estimate on the value of w ($= h$). Thus $w \leqq \ell$ implies that $5w \leqq \ell + 4w = 72$. Hence $0 \leqq w \leqq 72/5$ ($= 14.4$).

Now the conjecture that the students will want to prove is that $V(12) \leqq V(h)$, since most of them will have guessed by now from their graphs that 12 is the key value of h. One can emphasize here (as well as other places) the difference between hunch and proof. Both have an honorable place in mathematics. Both are equally valuable. It's like the starter and the transmission in a car. Each finds the other indispensable; rather, the driver finds each irreplaceable.

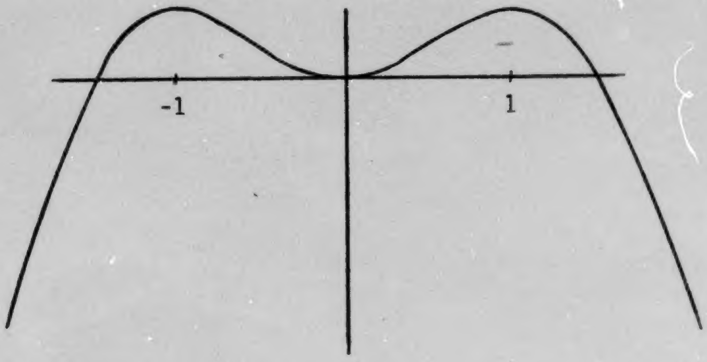
Now we consider the difference expressed in the conjectured inequality.

$$\begin{aligned}
&V(12 + h) - V(12)\\
&= 4[(18 - 12 - h)(12 + h)^2 - (18 - 12)\,12^2]\\
&= 4[(6 - h)(144 + 24h + h^2) - 6\cdot 12^2]\\
&= 4(-18h^2 - h^3) = -4h^2(h + 18) \leqq 0
\end{aligned}$$

But this last inequality is obviously true for all nonnegative h, and negative h are quite out of the running. Thus we conclude that the critical h which gives maximum package size is $h = 12$ and $\ell = 72 - 4(12) = 24$. Thus $V_{max} = 24(12)^2 = 3456$ cu. in. This proof should be understandable to any calculus student.

The following is another problem in this same genre of precalculus challenge that helps students to work with basic concepts and to strengthen their hold on algebraic tools.

Problem. Find high and low points of the function $y = 2x^2 - x^4$. Here one can use symmetry as a helpful tool and appeal directly to the purely algebraic method of completing the square. Once the graph of the function under consideration is drawn, students will have no trouble in putting forward a reasonable conjecture (reasonable as judged from a rough sketch) as follows: $y \leqq 1$. Then, as soon as they see that y is equal to 1 at x = 1 (and hence at x = -1 too), they're "home" so far as maximum values are concerned. One can also use this problem to cite the difference between relative and absolute extrema, since the origin proves by simple argument to be a relative minimum for y. See the graph below.



The ability to differentiate and integrate is much easier for students to pick up themselves than are the various techniques for interpreting physical situations and formulating them mathematically. From the student's point of view (if he could state it thus at this stage in his intellectual pilgrimage), this use of the instructor is the optimal solution to the student's (unexpressed) desire to get as much as he can from the college or university he has chosen. If the instructor emphasizes only those things that the student can readily learn on his own, it is a poor investment of time and effort.

Use of even and odd functions. Let f be a function which is defined on the interval $-a \leqq x \leqq a$. Then f is said to be even if $f(-x) = f(x)$ (thus the graph is symmetric about the y-axis) and f is odd if $f(-x) = -f(x)$ (thus the graph is symmetric about the origin).

What about the sum of even functions? Invite educated guesses. Push for reasons. Is the behavior like that of even numbers (sum of two evens is even)? How about proving that the sum of even functions is an even function? Go through the proof with the students:

$$(f + g)(-x) = f(-x) + g(-x) = f(x) + g(x) = (f + g)(x)$$

Further conjectures, with proofs in the spirit of the one just given, are suggested. Better still, try to get the students to do the conjecturing. (Pretty soon the word gets around to the other students that when they come into my class they are likely to get a "conjecture lecture.")

Conjecture. If f is even, then cf is even.

Conjecture. If f is even, then f + c is even.

Conjecture. The sum of odd functions is an odd function.

Conjecture. The product of two even functions is an even function.

Conjecture. The product of two odd functions is an even function.

Now ask the class what conclusion can be drawn about the uses of the words "odd" and "even" in the two cases, integers and functions. One can carry this even further by asking them to find some other mathematical (or nonmathematical) objects, together with an operation that connects them properly ("binary operation," without using the word) which behave in the same way as even and odd numbers under multiplication. If one has the time, this can be pushed into the direction of algebra and its isomorphisms.

Now start looking for functions that satisfy the definitions of even and odd. If the students can think of the first one, it's better; the others will then come more easily. Very likely, even powers of x will be suggested by them and then odd powers of x. The trigonometric functions are likely candidates too, unless they are not expected to take up the part of the course dealing with trigonometry until much later. Absolute value of x and fractional functions like $1/x^2$ are other examples that might occur to them.

The notion of bounded function. If the domain of f is $a \leqq x \leqq b$ and if there is a number M such that $|f(x)| \leqq M$, then M is a bound of f on its domain; we call f a bounded function and say that f is bounded by M. Now one can consider the familiar functions and determine whether they are bounded. Powers of x and all nonconstant polynomials are without bound. The function $1/(1 + x^2)$ is bounded. Let the students try to find a bound. Show how to estimate the values of this function. These are beginnings that will help students later on when they are supposed to think of the idea of estimating on their own and make some reasonable estimate of values of a function.

Conjecture. If f is bounded by M and g is bounded by N, then f + g is bounded by M + N.

Give the proof, which is another opportunity to emphasize what might still be a little novel to them: using the definition of $(f + g)(x)$. The proof essentially involves the triangle inequality. If students don't know the triangle inequality, you can give them hints about its origin, ask them to think about it, or cite it in the text.

Use of the function $f(x) = 2^x$. In contrast with e^x or 10^x, the function 2^x seems much easier to graph. Student's attempts at graphing it (or in assisting you to graph it) will provide valuable information about their knowledge or ignorance of exponents (negative and fractional). Sample computations for plotting the graph appear below:

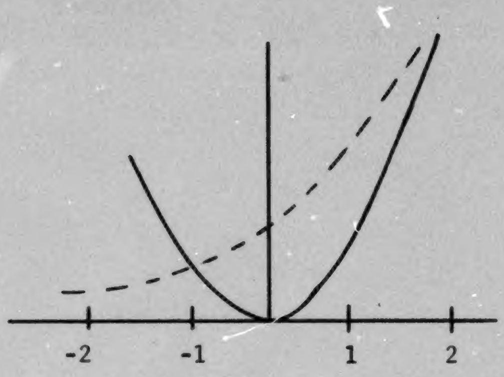
$2^{\frac{1}{2}} \cong 1.4$

$2^{\frac{3}{2}} \cong 2(1.4)$

$2^{-\frac{1}{2}} \cong .7$

$2^{\frac{1}{4}} \cong 1.2$

$2^{-\frac{3}{2}} \cong (.7)^3 \cong .34$



Once the graph is sketched, one can ask for the value of 2^x when x = 8/35, 9/10, etc. Draw out from the students the assumptions used when we connect the points on the graph. Remind them that little lines can sometimes result from big assumptions.

Recalling an earlier problem (find solutions of $x^2 = 2^x$), we can graph $y = x^2$ and $y = 2^x$ on the same set of axes. Then one solution is obvious. Let them puzzle over the solution in the second quadrant. You might want to make it a homework problem. Perhaps we'll return to the function 2^x and its graph later to extract further information regarding slopes.

The Relevance Problem. It is to your advantage to choose examples according to the interests of the class. Obtaining depth and breadth simultaneously is almost impossible. Unless the class is unusually homogeneous in its makeup, some will be dissatisfied by some of your choices. Since this is the main thrust of this conference--relevance of problems to actual situations, as well as the more obviously contrived ones--it is good to look further into the matter of relevance.

To Students: Make the calculus class relevant to your own field. I cannot possibly cover all the applications that are possible. Indeed, as the course progresses I shall be glad to get from you examples that you think are interesting and that you would like to share with the rest of us. Since I'm familiar with a number of these already (but far from all of them), I'll be glad to serve as a kind of filter or screening for them.

One way you can make the course relevant to your own special interest is to be alert in your other classes that might offer some real problem situations that calculus can be focused upon. Bring them back to this class for discussion and consideration, if you will.

To Teachers: It is a good idea to keep tabs on what other teachers are doing in courses that are thought by some, or are asserted by the catalogue, to depend upon the calculus course you are teaching. Check with the science teachers, in particular, to see how much mathematics they expect their students to know. Don't hesitate to point out to students the areas in which calculus will be used, especially when you are introducing certain ideas in precalculus problems like those discussed in the foregoing lectures and earlier in this lecture.

Term Papers. I have found the use of term papers to be very successful for most students. Assign term papers with the understanding that they can enhance their grade, but not hurt it. The amount of credit which you might want to give for a term paper (e.g., in lieu of a quiz, as an outside quiz, or whatever) is up to you as teacher. I tell them, "Write a five-page paper (or, more accurately, a certain number of words, since paper size varies) on the subject 'Calculus and _______________'." The blank can then be filled in by any subject in their range of interest, e.g., engineering, physics, optical illusions, ecology, simple harmonic motion, micro-economics, predator-prey relations. Ask the student to interview at least one faculty member in the process of his search for information. Have him ask why calculus is prerequisite or corequisite in other courses related to his topic. Ask what he expects freshmen in calculus to learn. Have him consult at least one book on his subject which is not a mathematical text. Encourage him to examine one or two problems and to write something that he finds significant about them.

Since some faculty members are touchy about being approached by students in this way, the students might be cautioned not to appear to ask any staff member to write something for him. He should try to get by on oral communication with the interviewee.

Some students will find this task tedious. Others will go wild with the idea and write three times as much as requested. They should all have a more realistic and wholesome respect for calculus as a result of their mini-research efforts and should all learn more about calculus, too.

In large classes you might pass the papers around, asking those interested in a given area to read papers of others in the same area. Each student who reads the paper of another student can then be requested to write a paragraph of comment on his treatment. And it will all mean a good source of problems for yourself as teacher! It should hopefully mean a closer tie-in with other departments on the campus, too.

CALCULUS

Lecture 4: Where Calculus Teaching Fails

Today we want to go trouble-shooting. We are looking for those topics or areas which seem to be especially troublesome for the student of the calculus. You have gotten underway and the sailing seems smooth enough. The student has his difficulties but nothing seems to be unusual. Then suddenly he is confronted by something so different from all his previous mathematical experiences that he is completely buffaloed. He sees you give a proof which he could not repeat if his life depended upon it. It just doesn't hang together with the rest of what he has presumably learned in the course. He doesn't know what to do about it, and it is a rare student who will approach his instructor with his trouble at this stage. The first day after he is exposed to this impossible proof, or whatever it is that troubles him, he is bewildered and dazed. If the difficulty isn't cleared up immediately, the mystery only deepens and he has less and less nerve to ask the instructor for an explanation of what's bothering him. It has become a psychological barrier to further absorption of calculus and is in danger of becoming a wedge to force him out of the course, dampen his enthusiasm for mathematics, and even prompt in him thoughts of dropping out. Since communication is often only one-way when the student is in the throes of such an experience, the teacher must be extra sensitive if the difficulty is to detected and treated. This is the point at which every calculus teacher (and every teacher, period!) must be a psychologist, at least a part-time psychologist, and on duty as such for each mathematics course he is teaching.

What are some of the problem areas that need special treatment because of the frequency of their occurrence? I want you to tell me how you feel about this. What are the areas that you are sensitive to by virtue of your experience, either as learner or as teacher? (At this point various problem areas were suggested by participants, as follows:

1. $\lim_{x \to 0} (\sin x)/x$	5. $(x^n)' = nx^{n-1}$
2. mathematical induction	6. integrability of continuous functions
3. $\lim_{x \to a} f(x) = f(a)$	7. chain rule for derivatives
4. derivative of a product	8. differentials)

There are three desiderata from algebra:

1. How to factor $a^n - b^n$ for arbitrary integral n. If you ask the students to do this one, don't be surprised at the great variety of stabs at the answer. If you really take the trouble to read them through, it will "stab your spirit broad awake" that they really don't know much about factoring.

2. Rationalize the numerator. This can be motivated by the fact (in your own mind) that it is the sort of thing needed a little later in formal differentiation and also in estimation problems involving difference or radicals. Thus

$$\sqrt{x+h} - \sqrt{x} = \frac{(\sqrt{x+h} - \sqrt{x})\ (\sqrt{x+h} + \sqrt{x})}{\sqrt{x+h} + \sqrt{x}} = \frac{h}{\sqrt{x+h} + \sqrt{x}} .$$

3. The quadratic formula. Actually very few of the students know this. You would probably not insult their intelligence at all if you simply proceeded to talk about it as if they had never seen it.

Derivative of the product of two functions. After the students have the definition of "f is differentiable at x_o," one can assign for homework some of the problems given in the appropriate sections. If it does not appear in the text (and it usually does not), one can find for them the derivative of f^2:

$$(f^2)(x) = [f(x)]^2$$

$$\frac{f^2(x+h) - f^2(x)}{h} = \frac{[f(x+h)]^2 - [f(x)]^2}{h}$$

$$= \frac{f(x+h) - f(x)}{h} [f(x+h) + f(x)]$$

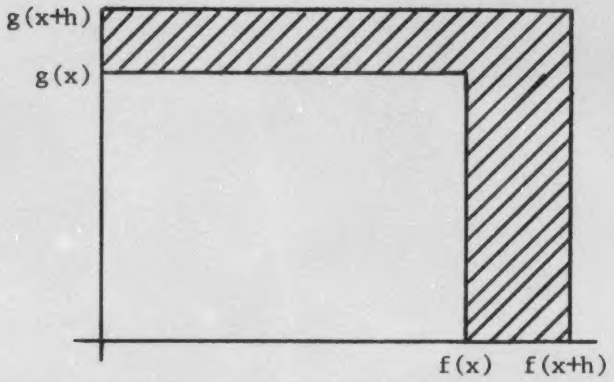
$$\longrightarrow f'(x)\ [2f(x)]$$

We can then reduce the problem of finding the derivative of the product to finding that of f^2 by means of the identity

$$fg = \tfrac{1}{2}[(f + g)^2 - f^2 - g^2].$$

Now there is another way of obtaining the formula for the derivative of a product, which uses a kinematic interpretation. The point here again is to get the student to feel good about his correct intuitions and to trust them. But somehow we must also try to dissbuse him of his false intuitions, if that is a proper term. The advantage of using the derivative of f^2 is that it can be easily generalized to obtain the derivative of f^n for integral n. The rational algebraic functions are a good category to work with in developing the rules of differentiation because they already are somewhat familiar to the student.

Kinematic interpretation of (fg)'. Assume that two men leave from the origin, one walking at a certain rate along the positive x-axis and the other walking upward along the positive y-axis. The land they are to be given in this land rush is that defined by the rectangle they determine (see the figure). Suppose that after x hours the first is located f(x) units from the origin on the x-axis and the second is located g(x) units from the origin on the y-axis. Now the problem is to determine the rate at which the rectangle of dimensions f(x) and g(x) is changing, with respect to the independent variable x.



Clearly, the difference $(fg)(x + h) - (fg)(x)$ is pictured as the difference between the areas of two rectangles ("inner" and "outer"), so we can express the area of the region representing this difference as the following sum:

$$[f(x + h) - f(x)]g(x + h) + [g(x + h) - g(x)]f(x)$$

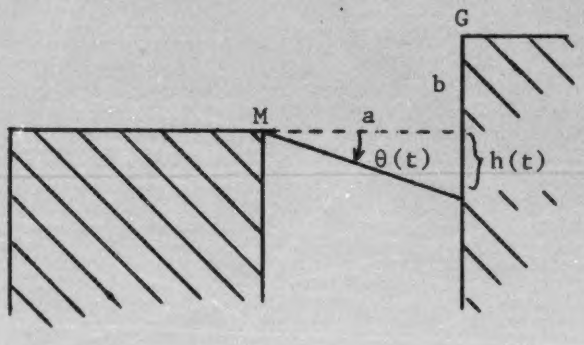
The analysis may now be completed in the usual way. The advantage of the kinematic illustration is that it seems reasonable and not at all tricky or artificial (once the idea of the problem is accepted). This leads readily to the usual formula for the derivative of fg:

$$\frac{f(x + h) - f(x)}{h}\, g(x + h) + \frac{g(x + h) - g(x)}{h}\, f(x)$$

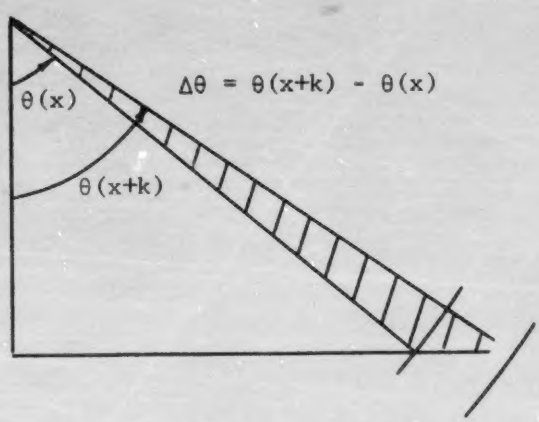
$$\longrightarrow f'(x)\, g(x) + g'(x)\, f(x)$$

Derivatives of the trigonometric functions. The case of the clumsy goat: On a hiking trip you come upon a chasm which separates you from a mountain goat who is located on a ledge 50 feet above you and on the edge of the canyon wall, which is 100 feet away. While you are looking, to your horrified amazement, he slips and falls straight down. Your eye follows him--all the way down. Later, in a more meditative mood, you recall the motion of your head as you watched him fall. It soon becomes apparent to you that your head was turning with a velocity that had some variation. Further cogitation suggests that in the process of turning your head downward, there might have been a maximum rate of turning. Then you begin to puzzle over it, and before you know it, you have a full-blown problem on your hands and are determined that if the difficulties are too great to solve mentally, you will work it out on paper upon returning to camp from your hike.

When you put the problem to your students and invite them to think about it, some are likely to answer that the rate of turning was most rapid when the goat was on a horizontal level with you. But you remind the student that at this point the goat has fallen for 50 feet and hence has built up considerable velocity. Some might offer the idea that the angle is changing most rapidly when the velocity is greatest, i.e., when he hits the ground. But then you point out that, unless the chasm is very shallow, your head is approaching asymptotically a position for which the eyes look straight down (we assume here that the line of vision is maintained by moving the whole head and that the eyes do not change their position within their sockets--thus it is head motion that we really mean when we talk about changing the direction of vision). After a few more questions and perhaps some further student guesses, they feel that it is a genuine problem (as opposed to a trivial one), and they are motivated to listen to a more careful and detailed analysis of it. If you substitute a for 100 and b for 50, the manipulation is not so bad. If h(t) is the height of the goat at time t, measured from the horizontal plane of your own position, then $h(t) = b - \frac{1}{2}gt^2$.



The next stage in the problem is the formulation of the angle of vision θ as measured from the horizontal. At first it is an angle of elevation. When the goat falls below horizontal eye-level, the angle is one of depression. Clearly, the definition relation is $\tan \theta = h(t)/a$. Thus our problem is now concerned with the derivative of θ with respect to t. Another diagram (to the right) explains our approach to finding this derivative. We catch the shaded region between sectors of two circles, as indicated by the arcs, and use the familiar formula for the area of a "piece of pie." This formula is unknown to some students, but you can derive it using a circle and the simple idea of proportionate area. This gives the inequality



$$\tfrac{1}{2}\left(\sqrt{1 + x^2}\right)^2 \Delta\theta < \tfrac{1}{2}k < \tfrac{1}{2}\left(\sqrt{1 + (x + k)^2}\right)^2 \Delta\theta,$$

leading, with a little manipulation, to the derivative of arctangent, which is what we seek. The last step is taking the limit as k approaches zero:

$$1 + x^2 < \frac{k}{\theta(x+k) - \theta(x)} < 1 + (x + k)^2,$$

so

$$1 + x^2 \leq \frac{1}{\theta'(x)} \leq 1 + x^2$$

Now the problem of the clumsy goat can be completed.

CALCULUS

Lecture 5: Miscellaneous Applications and Continuations

Completion of the Case of the Clumsy Goat. To simplify the problem, we take $a = 1$ and leave b arbitrary. Then the defining equations are $h(t) = b - \frac{1}{2}gt^2$ and $\theta(t) = \arctan h(t)$.

$$h'(t) = -gt \qquad h''(t) = -g$$

$$\theta'(t) = \frac{h'}{1+h^2} \qquad \theta''(t) = \frac{(1+h^2)h'' - 2h(h')^2}{(1+h^2)^2}$$

$$\begin{aligned}
\theta''(t) = 0 &\Longleftrightarrow (1+h^2)h'' - 2h(h')^2 = 0 \\
&\Longleftrightarrow (1+h^2)(-g) - 2hg^2t^2 = 0 \\
&\Longleftrightarrow h^2 + 2gt^2h + 1 = 0 \\
&\Longleftrightarrow 3h^2 - 4bh - 1 = 0 \\
&\Longleftrightarrow h = \frac{2b \pm \sqrt{4b^2 + 3}}{3}
\end{aligned}$$

Observe that g is missing, and hence the solution is independent of the constant of gravitational acceleration. That is, as long as the force of gravity is manifested in constant form, it does not affect the value of h.

Now we discuss the choice of sign. The plus sign clearly gives an h larger than $4b/3$ and therefore beyond the limitations of the physical problem that generated the initial equations. A brief explanation should convince the student that

$$h_b = \frac{2b - \sqrt{4b^2 + 3}}{3},$$

where we have indicated the dependence of h upon b by writing the latter as a subscript. Noticing that h_b is always negative, we can now formulate some idea of where the h_b occurs for the maximum rate of turning of the goat watcher.

There are two limiting cases to consider. When $b \to 0$, we have $h_b \to -\frac{\sqrt{3}}{3} \cong -.58$, whence $t \to \frac{1}{4}\sqrt{.58} \cong .19$. For the case $b \to \infty$, we have

$$\lim_{b\to\infty} \frac{2b - \sqrt{4b^2 + 3}}{3} = \lim_{b\to\infty} \frac{-3}{3(2b + \sqrt{4b^2 + 3})} = 0.$$

There is yet another answer that might occur to some of the students: at the point of impact with the ground, the goat watcher's head has suddenly stopped turning. At _that_ point, therefore, the rate of turning is the greatest. (Of course, if it is desired, the problem can be rephrased in order to eliminate this candidate for a solution.)

Composition of functions. We define $(f \circ g)(x) = f(g(x))$. The function $f \circ g$ is called the composite of f and g. We may probe the significance of this definition using, for instance, our previous discussion of even and odd functions. Proving certain simple results will help the students develop an understanding of proof and rigor. Some will immediately feel at home with these notions and will prefer making these types of investigations to "working with numbers." You can invite their response to questions embodied in conjectures such as these:

Conjecture. If f is even and g is odd, then $f \circ g$ is even.

Conjecture. If f is odd and g is even, then $f \circ g$ is even.

The proved conjectures may then be displayed in a table.

The chain rule will make more sense to the students when they understand composition of functions.

Monotonic Functions. To say that a function f is _increasing_ means that if $a < b$, then $f(a) \leqq f(b)$. Now we may discuss conjectures similar to those which we considered in our development of previous concepts. For example, if f and g are increasing, then $f + g$ and cf are also increasing. See what the students think about f^2, when f is increasing. Ask them about the simple special case $f(x) = x$. If they don't think to consider negative x, suggest it and try to get _them_ to draw the conclusion that $f^2(x) = x^2$ is not always increasing. Then one can ask whether there are circumstances under which the fact that f is increasing _does_ imply that f^2 is also increasing. "Little" situations like these should help them to cultivate care in proof and a reticence to jump at conclusions. The subject of decreasing functions can be assigned as homework.

Convexity. After students are well into the study of calculus, one can take up convexity. This concept can be illustrated with simple functions and easy-to-draw graphs. It will give them a new idea about objects already familiar.

Definition. A set S is convex if whenever points P and Q are in S, the segment PQ is also in S. A function is convex over $a \leqq x \leqq b$ if the region above its graph is convex. Give examples such as interiors of the circle, square, regular polyhedra, shadow of an ice cream cone (full), a balloon (with or without basket), a dirigible (without basket). Also give examples of non-convex sets such as a gnomon-like region, crescent, T-shape, hammer, sickle, swastika.

Definition. A function f is m-convex (midpoint convex) over $a \leqq x \leqq b$ if, for all points $x_1, x_2 \in [a,b]$,

$$\frac{f(x_1) + f(x_2)}{2} \geqq f\left(\frac{x_1 + x_2}{2}\right).$$

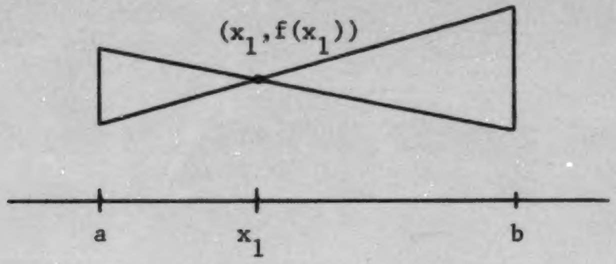
This is a two-point property which is amenable to the use of the computer. Students should be urged to make conjectures regarding the relationship between convexity and m-convexity.

Conjecture. If $a > 0$, then $f(x) = a^x$ is convex.

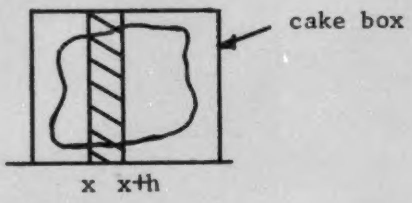
The proof of this conjecture requires showing that $\frac{1}{2}(a^s + a^t) \geqq a^{\frac{1}{2}(s+t)}$. This is easily done by squaring (which some of the students will doubtless think to do). It is also a restatement of the fact that the geometric mean of two numbers is not greater than the arithmetic mean.

Convexity can be related to the calculus in several ways. One immediately rewarding way is related to concavity. The definition of convexity for a differentiable function provides the second derivative test for upward concavity, for example.

Lipschitz Condition. Another rather simple geometric idea for the calculus student is the k-Lipschitz condition. A function f has the k-Lipschitz property if for any pair of points x_1, x_2 in the domain $a \leqq x \leqq b$ of f, we have $|f(x_1) - f(x_2)| \leq k\,|x_1 - x_2|$. Geometrically, this has a rather simple interpretation. It means that if we fix a point x_1, then the graph of f must lie in a region which has the shape indicated in the diagram.



Let us return to the problem of halving a cake. Again we use a cake from the Irregular Bakery, and we define L(x) to be the amount of cake to the left of a vertical line at x. We estimate the area of the cake contained in a vertical strip of width h and immediately to the right of the vertical line through x. We have



$$0 \leq |L(x + h) - L(x)| \leq k|h|,$$

so

$$0 \leq \lim_{h\to 0} [L(x + h) - L(x)] \leq 0.$$

This means that $\lim_{h\to 0} L(x + h) = L(x)$, and hence the cake-half function L(x) has the property of smoothness (if one continues to shy away from the word "continuity").

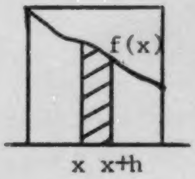
Another result is a proof of the Fundamental Theorem of Calculus.

Proof: If f is a decreasing function which is k-Lipschitz for some positive k and if L is interpreted as before,

then we must have $hf(x + h) \leq L(x + h) - L(x) \leq hf(x)$,

so that $f(x + h) \leq \frac{1}{h}[L(x + h) - L(x)] \leq f(x)$.

Since f is k-Lipschitz, we have $\lim_{h\to 0} f(x + h) = f(x)$.



From this fact and the above inequality, it follows that $L'(x) = f(x)$. This is one form of the Fundamental Theorem of Calculus.

One can also increase students' familiarity with the k-Lipschitz condition by posing certain questions in conjecture form.

Conjecture. The product of two k-Lipschitz functions is not necessarily k-Lipschitz.

For example, x is 1-Lipschitz, but x^2 is not 1-Lipschitz--nor is it k-Lipschitz for any possible k. The function simply grows faster than any constant multiple of $ax + b$.

Conjecture. Every function that is convex is also continuous.

This one is false. The interior points give no trouble, but the boundary points *can*. The function f defined below is convex, but not continuous.

$$f(x) = 1, \quad x = 0$$

$$= x^2, \quad x > 0$$

$$S_f = \{(x,y): x \geqq 0,\ y > f(x)\} \text{ is convex.}$$

Sphere-cutting with a knife have two (parallel) blades. Given a knife with two parallel blades, where should we make the incision on a spherical cake in order to maximize the volume of the slice, the slice being defined as that portion of the cake cut between the blades? Most students will recognize the geometry of this problem and decide to cut so that both blades are the same distance from the polar plane having given normal.

Where should the cut be made to obtain a minimum volume for the slice, assuming each blade at least touches the cake? (There is no difficulty here, either.) Suppose we want to maximize the icing; then where do we cut? The answer turns out to be independent of the cut, provided each blade contacts the cake. Unless the students have had solid geometry, they will probably not know this result. [See Ogilvy, C. Stanley. *A Calculus Notebook*.]

The ring problem. For this application, consider a ring which was cut from a sphere by removing the material within a right circular cylinder. The remaining band is 1/8 inch wide. The problem is to show that if the band-width of 1/8 inch is fixed, then the amount of material in the ring (i.e., its volume) is independent of the radius. Even though the ring is made from a sphere the size of the earth, this is still true. They might have difficulty believing this one!

CALCULUS

Lecture 6: An Arch, A Cup, and e

I. The St. Louis Arch--"Gateway to the West"

For your class in calculus, the St. Louis arch can become the gateway to an interesting, if sometimes difficult, set of applications of the calculus. The arch is 630 feet high; its outside width at the base is also 630 feet. The cross-section of the arch normal to the outside curve is an equilateral triangle. An arch meeting these specifications could have several possible forms. There are many questions that can be raised about the arch, and some of them will make good problems.

1. In a more general context, you can ask the class to name curves for which each of the following can be easily determined: arc length, area between the curve and the x-axis, volume of the region obtained by rotating the curve about the x-axis or the y-axis, surface area of the region obtained by rotating the curve about the x-axis or the y-axis. The circle and the straight line will probably occur to your class. But I know of only one other curve with this kind of "simplicity." After they guess a bit, if they do not hit upon it, suggest the catenary.

2. The sections of the arch were put in place from the ground up. How high did they reach before it was necessary to prop the structure to prevent it from toppling over, assuming (falsely) that it was not anchored from below?

3. Assuming that the wind has a fixed velocity and that its force upon the arch is proportional to the amount of area exposed (or whatever the assumption should be, as suggested by the aeronautical engineers), from what direction should the wind come in order to maximize its force against the median plane of the arch?

4. Determine whether the inside curve of the arch is a catenary. If it is, then the two catenaries, inside and outside, easily determine the cross-section of the arch. In any case--in whatever way the arch is formed, what is the amount of steel of thickness t inches which is required to make the arch? Also, find the volume enclosed. Area pertains directly to the cost, volume to the problem of air-conditioning the interior.

5. We may formulate a problem about the classroom model rather than the actual arch: if we rotate the arch by small displacements from the position of equilibrium and in the median plane, describe mathematically the resulting vibratory motion.

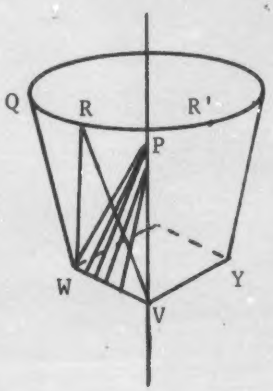
(Some of the above problems are purposely left undefined. The missing links in information can either be presented to the class or left for the class to determine for themselves.)

II. The Air France Cup

The plastic cup used by some airlines, such as Air France, has a circular top, a square base, and dimensions such that the cups are stackable. The edge of the square and the radius of the circular rim are both about $1\frac{1}{2}$ inches.

Now suppose we know only that the cups are stackable, have circular rim and square base, and that the set of points interior and on the cup (neglecting its thickness) represents the smallest convex set that so conforms. Determine the shape of the cup.

First Method. Consider the diagram. Take a point P on the circular rim nearest a vertex V of the square base. Since PV lies on the surface, as does WV, the plane PVW must lie on one side of the cup. Similarly, the plane PVY lies on one side of the cup. This is a direct consequence of the definition of convex set.



The local tin-smith or air-conditioning expert will probably recognize this problem as an old one, because it is not uncommon for makers of ducts to have to go from a round pipe to a square orifice or vice versa.

Clearly, if Q is to W as P is to V, and if R is the midpoint of the arc $\overset{\frown}{PQ}$, then RWV is the plane tilted farthest away from the axis α of the cup, it being assumed that the plane and the cup have a positive area in common. Thus RWV is contained entirely on the surface. As R moves toward P, the plane it generates with WV lies inside the cup until P is reached. Thus PV is on the cup. Continuing in this way, we determine the shape of the Air France cup.

Second Method. Take the vertex V of the base and draw PV. Then V$\overset{\frown}{RP}$ is on the surface, and so is V$\overset{\frown}{R'P}$, where $\overset{\frown}{RR'}$ is a quarter circular arc. Once R' is reached, the plane R'VY is on the surface, etc. This technique, continued again, determines the same Air France cup.

The next problem is to find the area of the cup. This is accomplished by dissecting the cup and finding areas of simpler figures. The cup can be decomposed into a central square pyramid, four peripheral conical pieces (use Cavalieri's theorem to find their volume), and four tetrahedral pieces, all of whose dimensions can easily be computed.

Finally, you can ask for the volume of the cup. In the light of the foregoing decomposition, the volume may be easily computed from the nine figures into which the cup was broken.

III. On e

The number e appears in several different contexts in calculus. There is seldom any unity in the treatment of such occurrences and there is likely to be a little mystery, at least, for some students.

1. First definition. Graph 2^x, 3^x, a^x. Then observe that since $\lim\limits_{h\to 0} \frac{a^{x+h} - a^x}{h} = a^x \lim\limits_{h\to 0} \frac{a^h - 1}{h}$, it must be the case that if the function $y = a^x$ has a derivative at $x = 0$, then for each real x it also has a derivative: a^x multiplied by the derivative at $x = 0$. This gives the functional equation $f'(x) = f'(0)\, f(x)$. Then we define e to be that base a for which

$\lim_{h\to 0} \frac{a^h - 1}{h} = 1.$

Now how does one find the limit? To determine the slope of 2^x at $x = 0$, one can estimate the slope on the right and the slope on the left at $x = 0$. If this is done for the intervals $(0,1)$, $(0,\frac{1}{2})$, ..., $(0,1/n)$, it can be found that $\frac{1}{1/n}(2^{1/n} - 1) = n(2^{1/n} - 1) \to .7$. More generally, we define

$L(a) = \lim_{n\to\infty} \frac{a^{1/n} - 1}{1/n}$. We follow a similar process to determine the estimation of the slope on the left of $x = 0$. This leads to the estimate .7 for the left-hand slope. These two, the right-hand and the left-hand estimates, are actually closely related, since $n(a^{1/n} - 1) = n(1 - a^{-1/n})\cdot a^{1/n}$. This means that if we can show that $a^{1/n}$ approaches 1 as n tends toward infinity, we have the existence of the derivative of a^x at $x = 0$ without further ado.

(One way out of the bind is to use Bernoulli's inequality, which the students should be exposed to anyway, and which might have already been presented):

Lemma. If $b > 1$, then $b^{1/n} \to 1$ as $n \to \infty$.

Proof: Assume that $b^{1/n} \to 1 + h > 1$. Then $b^{1/n} > 1 + h/2$ for all sufficiently large n. Then $b > (1 + h/2)^n \geq 1 + nh/2$, which can be made arbitrarily large. But this is impossible for a given constant $b > 1$. So $h \leq 0$. But $h < 0$ is even less realistic. So $h = 0$. Hence $b^{1/n} \to 1$ as $n \to \infty$.

2. Second definition. Set $\ln x = \int_1^x \frac{dt}{t}$, where $x > 0$. Then $(\ln x)' = 1/x > 0$. Define e to be that x such that $\int_1^e \frac{dt}{t} = 1$. We can estimate $\ln 2 = .7$; $\ln 4 = \int_1^2 \frac{dt}{t} + \int_2^4 \frac{dt}{t} \cong .7 + .7$. Thus $\ln 4 > 1 > \ln 2$. More generally, $\int_a^{ax} \frac{dt}{t} = \int_1^x \frac{ds}{s}$, where $at = s$. This magnification of scale by a (constant) can be made plausible by reference to sets that are enlarged or contracted by a single factor (depending upon whether $a > 1$ or $a < 1$, respectively). Hence we obtain the basic property for logarithms:

$$\ln ab = \int_1^{ab} \frac{dt}{t} = \int_1^a \frac{dt}{t} + \int_a^{ab} \frac{dt}{t} = \int_1^a \frac{dt}{t} + \int_1^b \frac{dt}{t} = \ln a + \ln b$$

This approach to e is clearly based upon a prior development of the logarithm function.

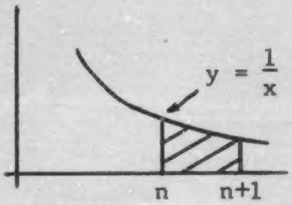
3. Third definition. This is the compound interest approach. Ask the students how money grows when it is deposited in a savings account. Gradually move toward the idea of a continuous compounding of interest. This easily leads to some symbol for the limit that appears in this compound interest problem. We define

$$e = \lim_{n\to\infty} (1 + 1/n)^n .$$

Assuming that we can interchange the order of taking a logarithm and taking a limit, we get $\ln e = \lim_{n\to\infty} n[\ln (n + 1) - \ln n]$. Now use l'Hôpital's rule to justify writing

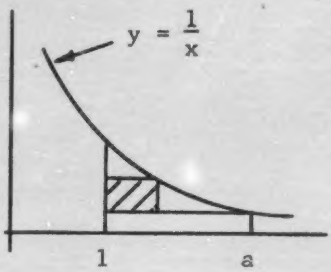
$$\lim_{x\to\infty} \ln (1 + 1/x)^x = \lim_{x\to\infty} \frac{\ln (1 + 1/x)}{1/x} = \lim_{x\to\infty} \frac{\frac{1}{1 + 1/x} (-1/x^2)}{-1/x^2} = 1.$$

Then we estimate the quantity $\ln (n + 1) - \ln n$ above and below, using circumscribed and inscribed rectangles.



This yields $\frac{1}{n + 1} < \ln (n + 1) - \ln n < 1/n$, from which we conclude $\lim \ln \left(\frac{n+1}{n}\right) = 0$. Now we find an exponential curve $y = a^x$ which passes through (0,1) and (b,c). Thus $c = a^b$. Take successively $b = \frac{1}{2}, \frac{1}{3}, \ldots$ and define a by $1 + 1/n = a^{1/n}$, whence $a = (1 + 1/n)^n$.

Returning for a moment to an earlier problem, we ask the maximum size for a rectangle placed above the rectangle R and below the curve $y = 1/x$.



We get $a^{\frac{1}{2}}$ as the abscissa of the right end of the approximating rectangle. The problem can be generalized so that if we take n - 1 points of subdivision between 1 and a, then the points that maximize the total area of the corresponding rectangles are the multiple powers of

$a^{1/n}$: these are $a^{1/n}$, $a^{2/n}$, $a^{3/n}$, ..., $a^{(n-1)/n}$.

Another technique for estimating the area of the region under the curve $y = 1/x$ and between 1 and a is to use $\frac{1}{x^{1+1/n}}$ below and $\frac{1}{x^{1-1/n}}$ above. Thus

$$\int_1^a \frac{dx}{x^{1+1/n}} < \int_1^a \frac{dx}{x} < \int_1^a \frac{dx}{x^{1-1/n}}$$

and

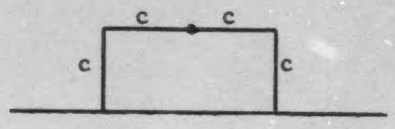
$$n(1 - a^{-1/n}) = - n\, x^{-1/n} \Big|_1^a < \ln a < n x^{1/n} \Big|_1^a = n(a^{1/n} - 1).$$

It is instructive for the student to see material on e from several points of view and at various times. Repetition is not bad, especially when it is not going in a circle, but rising on a spiral. Whatever the approach to e, it can be the occasion for the student in his own mind to bring together several strands of the subject and to begin his own process of mental braiding.

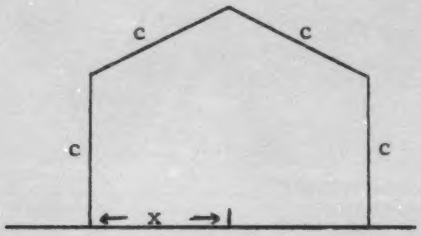
CALCULUS

Lecture 7: The Playpark Problem

A certain manufacturer makes a playpen whose flexible construction permits the linkage of its four sides (normally a square) to be attached to a wall (the side of a house, for example). They advertise that such an arrangement could easily double the child's play area.

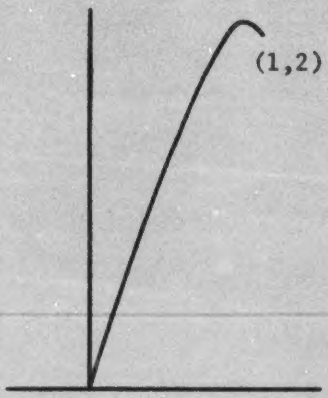


The problem is to determine whether there is a larger play area when the straight wall is allowed as the fifth side and the attachment to the wall is by means of right-angle braces.



In this case the area A is a function of the half-length x of the wall portion, $A(x) = 2cx + x\sqrt{c^2 - x^2}$. It is easily verified that $x_1 = c\sqrt[4]{\frac{3}{4}}$ gives the maximum value of A. The graph of $y = A(x)$ is itself interesting, as part of the determination of the maximum A. We have, for $c = 1$,

x	0	.5	.8	.9	1
A(x)	0	1.43	2.08	2.19	2



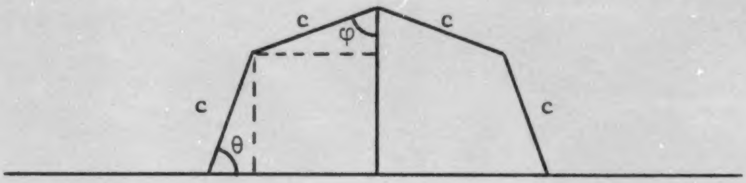
$$A(x_1) = 2x_1 + x_1\sqrt{1 - x_1^2} \cong 1.86 + .93\sqrt{.13}$$

$$\cong 2.196$$

Thus the maximum area is about 20 per cent larger than the two square units suggested by the manufacturer. But the restriction of the right-angle brace prevents a possibly larger area, so let us remove this restriction and consider the more general problem.

We let the braces be adjustable to any desired angle. By setting up the problem in terms of angles θ and φ as shown in the diagram below, we take advantage of symmetry. For the area we have

$$A(\theta,\varphi) = 2[c \sin \theta)(c \sin \varphi) + \tfrac{1}{2}(c \cos \theta)(c \sin \theta) + \tfrac{1}{2}(c \cos \varphi)(c \sin \varphi)],$$



whence θ and φ appear symmetrically. This suggests that θ = φ will satisfy the partial differential equations which must hold in order for A to have an extremum. We get

$$A_\theta = c^2(2\cos\theta\sin\varphi + \cos^2\theta - \sin^2\theta) = 0$$

$$\Longleftrightarrow -2\cos\theta\sin\varphi = \cos 2\theta$$

$$A_\varphi = c^2(2\cos\varphi\sin\theta + \cos^2\varphi - \sin^2\varphi) = 0$$

$$\Longleftrightarrow -2\cos\varphi\sin\theta = \cos 2\varphi.$$

Then $\theta = \varphi$ implies that $-\sin 2\theta = \cos 2\theta$. This occurs only in the second quadrant (whence $\theta = 3\pi/8$) and in the third quadrant (whence $\theta = 5\pi/8$). If we take the other angles of the triangle into consideration, we see that the solution $\theta = 5\pi/8$ cannot occur. The solution $\theta = 3\pi/8$ leads to the conclusion that the area is half an octagon's interior.

Another way to approach the problem of the adjustable braces is in terms of two degrees of freedom with one of them defined as a function of the other. We consider the classical 4-bar linkage problem with 3 bars of equal length. To begin more generally, we want to maximize the area of the quadrilateral having sides of length a, b, c, and d. Referring to the diagram below, we treat φ as a function of θ. Then the area A is given by

$$A(\theta,\varphi(\theta)) = \tfrac{1}{2}ab\sin\theta + \tfrac{1}{2}cd\sin\varphi(\theta),$$

where

$$a^2 + b^2 - 2ab\cos\theta = c^2 + d^2 - 2cd\cos\varphi(\theta).$$

So

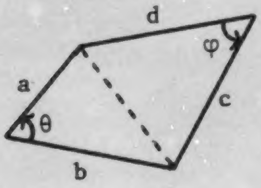
$$2ab\sin\theta = (2cd\sin\varphi)\,\varphi'(\theta), \text{ whence } \varphi'(\theta) = \frac{ab\sin\theta}{cd\sin\varphi},$$

and

$$\tfrac{1}{2}ab\cos\theta + (\tfrac{1}{2}cd\cos\varphi)\frac{ab\sin\theta}{cd\sin\varphi} = 0.$$

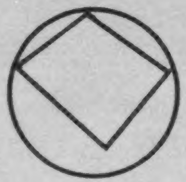
So

$$\cos\theta + \cos\varphi\frac{\sin\theta}{\sin\varphi} = 0, \quad \text{or} \quad \sin(\varphi + \theta) = 0.$$



This means that the quadrilateral can be inscribed in a circle. Note that the solution is independent of the lengths of the sides. In particular, we did not have to consider the special case of three sides' having the same length.

Here is a good place, if you like, to discuss the isoperimetric problem. The term will be new to the students, but the idea will seem reasonable now that they have done some extremal problems. That the circle has to be the solution for a curve of given length to form a region with maximum area will seem to many intuitively clear. It is easy to see that the region must be convex. If the proposed solution is not a circle, then it must have four points which are not on a circle. Consider the largest circle among those circles which can be constructed taking these four points three at a time. The vertex omitted must lie inside this circle; hence, the proposed region could have been enlarged. So the solution must be a circle.



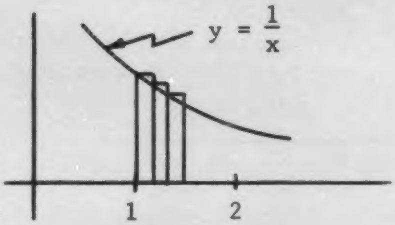
Integration of Decreasing Functions. As a rule we prefer to take left-hand endpoints and equal subdivisions in order to define approximating areas. Our basic definition is that of areal.

Definition. The areal of a function on the interval [a,b] is the limit

$$A_a^b(f) = \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(a + i\frac{b-a}{n}\right)\cdot\left(\frac{b-a}{n}\right).$$

More precisely, if we call this the left areal, then we can define similarly the right areal. First we apply the notion to the simple curve $y = 1/x$ for $1 \leqq x \leqq 2$. If the area under $1/x$ is A, then we set

$$\left(A_1^2\right)_n\left(\frac{1}{x}\right) = \frac{1}{n}\left(\frac{n}{n} + \frac{n}{n+1} + \dots + \frac{n}{2n+1}\right),$$



so that

$$\lim_{n\to\infty} \left(A_1^2\right)_n\left(\frac{1}{x}\right) = A_1^2\left(\frac{1}{x}\right),$$

the left areal. Thus, if f is decreasing, then the areal always exists and we have for the estimate of error:

$$0 \leqq \left|A_a^b(f) - \left(A_a^b\right)_n(f)\right| \leqq \frac{1}{n}[f(a) - f(b)].$$

Notice that here we can determine the area without knowing either the upper sum or the lower sum.

A similar sum can be set up for 1/x between 3 and 6. (The key is that 1:2 :: 3:6.)

In this context it also seems worthwhile to discuss the proof of the Fundamental Theorem of Calculus. If the function f is decreasing, we can begin with the Lipschitz condition

$$0 \leqq \frac{L(x + h) - L(x)}{h} \leqq k .$$

Then

$$f(x + h) \leqq \frac{L(x + h) - L(x)}{h} \leqq f(x).$$

Now we let h approach 0 in order to squeeze the middle quantity toward f(x). We raise here the question, too, whether the areal is an additive function on the line. That is, does

$$A_a^c(f) = A_a^b(f) + A_b^c(f)?$$

For a counterexample to the equality, we take

$$f(x) = \begin{cases} 1, \text{ if } x \text{ has a terminating decimal} \\ 0, \text{ otherwise.} \end{cases}$$

Now we observe that between any two terminating decimals there is a nonterminating decimal and, vice versa, between any two nonterminating decimals there is a terminating decimal. Recalling the definition of the areal for f, we see that

$$A_0^{\sqrt{2}}(f) = 0, \quad A_0^1(f) = 1, \quad \text{and} \quad A_1^{\sqrt{2}}(f) = 0,$$

whence

$$A_0^{\sqrt{2}}(f) \neq A_0^1(f) + A_1^{\sqrt{2}}(f).$$

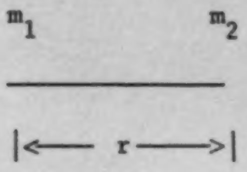
This means that the areal is not the same as the Riemann integral. The areal can exist without the Riemann integral's existing, but if the Riemann integral exists, then the areal must exist and be equal to it.

CALCULUS

Lecture 8: Newton's Law of Gravitation

Every particle in the universe attracts every other particle with a force which is directly proportional to the product of the masses of the particles and inversely proportional to the distance between them.

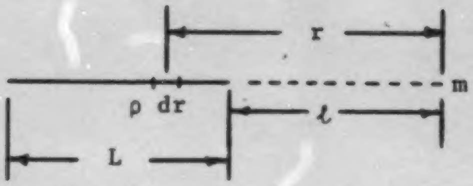
We are interested only in the magnitude of the force, which is a vector, so we write $F = G\frac{m_1 m_2}{r^2}$.



Examples:

1. Gravitational attraction of a homogeneous straight wire on a particle collinear with the wire.

$$dF = G\frac{m\rho\ dr}{r^2}$$



$$F = Gm\rho \int_{\ell}^{\ell+L} \frac{dr}{r^2}$$

$$= G\frac{m\rho L}{\ell(\ell + L)} = G\frac{mM}{\ell(\ell + L)}$$

Particle m "sees" the wire as a particle of mass M (= $L\rho$) located at the geometric mean of the endpoints of the wire.

2. Gravitational attraction of a homogeneous straight wire on a particle lying on the perpendicular bisector of the wire.

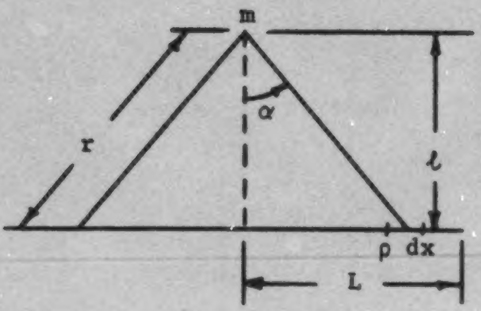
Because of symmetry, the x-components of F cancel and only the y-components need be considered. Two pieces of wire act symmetrically on m, so

$$dF = 2G\frac{m\rho\ dx}{r^2}\cos\alpha$$

$$= 2Gm\rho\ell\frac{dx}{(x^2 + \ell^2)^{3/2}}$$

$$F = 2Gm\rho\ell\int_0^L \frac{dx}{(x^2 + \ell^2)^{3/2}}$$

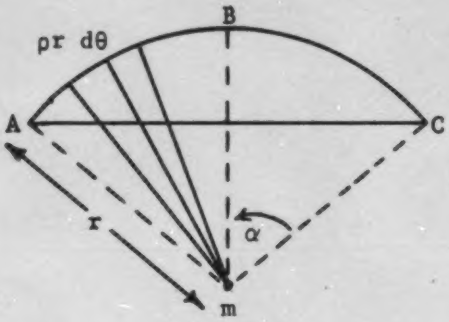
$$= G\frac{m(2\rho L)}{\ell\sqrt{\ell^2 + L^2}} = G\frac{mM}{\ell\sqrt{\ell^2 + L^2}}$$



Particle m "sees" the wire as though it were a point with mass M (= 2ρL) at a distance which is the geometric mean of ℓ and the distance to the end of the wire. As $L \to \infty$, $F \to \frac{2Gm\rho}{\ell}$; that is, the gravitational attraction to an infinitely long wire is inversely proportional to ℓ, not to ℓ^2.

3. Gravitational attraction of a homogeneous circular arc on a particle at its center.

Again, horizontal components cancel and only vertical components need to be considered.



$$dF = 2G\frac{m\rho r\ d\theta}{r^2}\cos\theta$$

$$F = 2\frac{Gm\rho}{r}\int_0^\alpha \cos\theta\ d\theta = G\frac{m(2m\rho\ \sin\alpha)}{r}$$

The length of the chord is $2r \sin \alpha$ and the mass of the chord is $2\rho r \sin \alpha = M$. Then

$$F = G\frac{m(2\rho \sin \alpha)}{r} = G\frac{m(2\rho \sin \alpha)r}{r^2} = G\frac{mM}{r^2},$$

and the particle "sees" the arc ABC as a particle with mass $2\rho r \sin \alpha$ at a distance r (say at B). Note that if $\alpha = 2\pi$, then $F = 0$ and all components of gravitational attraction cancel.

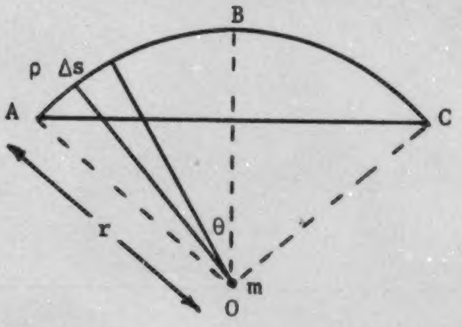
Another derivation does not use calculus but is indicative of the clever methods which can be used to solve many such problems.

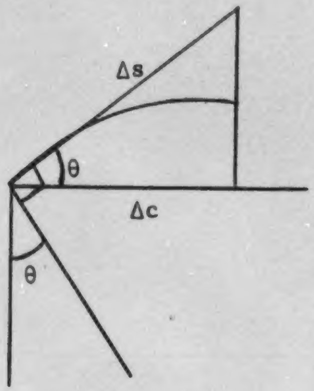
Again, the horizontal components cancel and we have to consider only the components parallel to OB. The component of attraction along OB of $\rho\, \Delta s$ on m at O is

$$dF = Gm\rho\, \Delta s \cos \theta / r^2.$$

If Δc is the projection on AC of Δs, then

$$\Delta c = \Delta s \cos \theta.$$





The mass of a wire of length Δc is $\rho\, \Delta c$, and $\rho\, \Delta c = \rho\, \Delta s \cos \theta$. If this piece of chord were moved to point B, its attraction on m at O would be

$$dF = G\frac{m(\rho\, \Delta s \cos \theta)}{r^2},$$

i.e., the same as the piece of circular wire of length Δs. So, if every particle of the chord AC were put at B, the total attraction on m at O would be the same as the total attraction of the arc ABC. The total mass of the chord AC is $2\rho r \sin \alpha$, so the total gravitational attraction of arc ABC on mass m at O is

$$F = 2G \frac{m\rho \sin \alpha}{r},$$

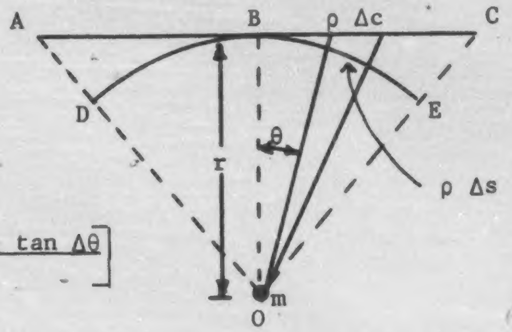
as before.

4. Gravitational attraction of a homogeneous straight wire on a point.

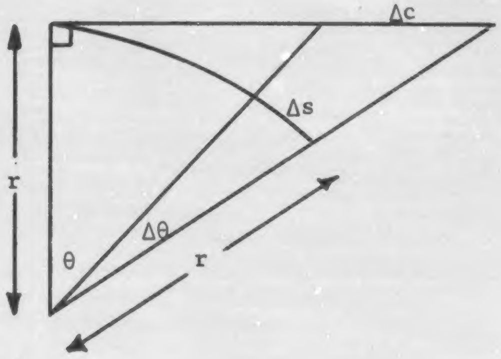
Draw a circle with center at O, tangent to ABC at B. The force of attraction of $\rho\, \Delta c$ and $\rho\, \Delta s$ on m at O are equal, because, as is shown below,

$$\Delta s \cong \Delta c \cos^2 \theta;$$

for, from the sketch,



$$\begin{aligned}
\Delta c &= r[\tan(\theta + \Delta\theta) - \tan\theta] \\
&= r\left[\frac{\tan\theta + \tan\Delta\theta}{1 - \tan\theta \tan\Delta\theta} - \tan\theta\right] \\
&= r\left[\frac{\tan\theta + \tan\Delta\theta - \tan\theta + \tan^2\theta \tan\Delta\theta}{1 - \tan\theta \tan\Delta\theta}\right] \\
&= r \tan\Delta\theta \left[\frac{1 + \tan^2\theta}{1 - \tan\theta \tan\Delta\theta}\right] \\
&\cong r\, \Delta\theta \left[\frac{1 + \tan^2\theta}{1 - \Delta\theta \tan\theta}\right] \\
&\cong r\, \Delta\theta (1 + \tan^2\theta) \\
&= r\, \Delta\theta \sec^2\theta.
\end{aligned}$$



Thus,

$$r\, \Delta\theta \cong \Delta c \cos^2 \theta;$$

and, from $\Delta s = r\, \Delta\theta$,

$$\Delta s \cong \Delta c \cos^2 \theta.$$

The distance from 0 to Δc is $r \sec \theta$. The attraction of $\rho\ \Delta c$ on m at 0 is

$$dF = G\frac{m(\rho\ \Delta c)}{(r \sec \theta)^2}$$

$$= G\frac{m\rho\ \Delta c \cos^2 \theta}{r^2}$$

$$\cong G\frac{m\rho\ \Delta s}{r^2}$$

(because, from above, $\Delta c \cos^2 \theta \cong \Delta s$); this is seen to be the same as the attraction of $\rho\ \Delta s$ on m at 0.

Then, because each particle of the wire has the same gravitational attraction on m at 0 as the corresponding particle of the circular arc, both in magnitude and direction, the total attraction of the wire is the same as the total of the circular arc. Let the angle subtended by the wire at 0 be 2α. The resultant attraction bisects this angle and, by the previous result,

$$F = 2G\frac{m\rho \sin \alpha}{r}.$$

Again, as the line increases in length, α increases and becomes $\pi/2$ for a rod of infinite length--so for a line of infinite length,

$$F = \frac{2G\rho m}{r},$$

as before (the previous result had ℓ in place of r).

5. The gravitational attraction of a homogeneous circular plate at a point on the axis of symmetry.

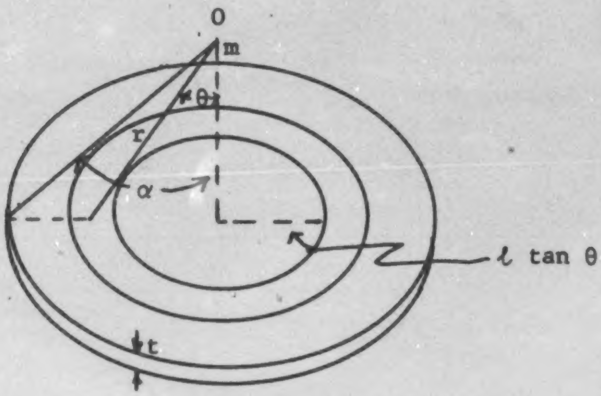
Consider the attraction of an annulus on a mass m at 0 (assume that all points of the annulus are at the same distance from 0--this avoids double integration).

$$dF = \frac{2\pi\ell \tan\theta\ \ell \sec^2\theta\ t\ d\theta}{(\ell \sec\theta)^2} m\rho \cos\theta$$

$$F = 2\pi G m\rho t \int_0^{\alpha} \tan\theta \cos\theta\ d\theta$$

$$= 2\pi G m\rho t \int_0^{\alpha} \sin\theta\ d\theta$$

$$= 2\pi G m\rho t\ (1 - \cos\alpha).$$



For an infinite plate, $\alpha = \frac{\pi}{2}$, and then

$$F = 2\pi m\rho G t, \text{ independent of } \ell.$$

If, as was believed until a few hundred years ago, the earth is an infinite flat surface, the acceleration of gravity would be constant, independent of height, according to Newton's Second Law, which in this case becomes

$$F = m(2\pi\rho G t)\ .$$

Comparing this with

$$F = ma,$$

we see that $a = 2\pi\rho Gt$, which is independent of the height above the surface of the earth.

Building models; particles. In considering motion and forces between the earth and the sun, we can regard each as a point mass. The diameter of the sun is 860,000 miles. The distance from the earth to the sun is 93,000,000 miles, about 100 times the diameter of the sun and 10,000 times the diameter of the earth. But, in considering the earth-moon system, we have to consider the earth and moon as spheres, for otherwise we cannot account for eclipses of the sun and moon; rigid spheres are acceptable here as models for the earth and moon. However, to study earthquakes the model of a rigid sphere is not appropriate; in that situation we consider the earth as a sphere composed of elastic material. Notice how the model for the earth may be simple or relatively complicated depending upon the problem being considered.

Newton needed the result of example 6 (below) to establish the theory of gravitation applied to the moon. The moon is close enough to the earth so that neither body can be considered as a point, so Newton had to know how to determine the attraction between two spheres. Supposedly, it took him 20 years to establish this result; publication of his theory of gravitation was held up for that length of time, until he had determined that a homogeneous sphere attracts an external particle as though the mass of the sphere were concentrated at the center.

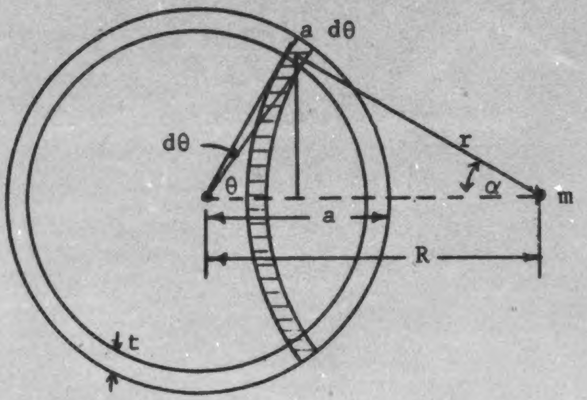
6. Gravitational attraction of a homogeneous spherical shell on an external particle of mass m.

Again, we need consider only the component of force parallel to $\overline{OP}$. Divide the spherical shell into rings. The mass of each ring is

$$\begin{aligned} dM &= \rho t\ 2\pi a \sin\theta\ a\ d\theta \\ &= 2\pi a^2 t\rho \sin\theta\ d\theta\ , \end{aligned}$$

and the component of gravitational attraction parallel to OP is

$$dF = Gm\frac{2\pi a^2 t\rho \sin\theta\ d\theta}{r^2} \cos\alpha.$$



We need to express cos α and r^2 in terms of θ. From the sketch,

$$\cos\alpha = \frac{R - a\cos\theta}{r},$$

and, by the law of cosines,

$$r^2 = a^2 + R^2 - 2aR\cos\theta.$$

Thus,

$$2r\,dr = 2aR\sin\theta\,d\theta$$

and

$$a\sin\theta\,d\theta = \frac{r\,dr}{R}.$$

Now write dF in the form

$$dF = Gm\frac{2\pi at\rho(a\sin\theta\,d\theta)}{r^2}\,\frac{R - a\cos\theta}{r}$$

$$= Gm\frac{2\pi at\rho r dr}{Rr^2}\,\frac{R - a\cos\theta}{r}\ .$$

To each side of the equation

$$r^2 = a^2 + R^2 - 2aR\cos\theta$$

add $R^2 - a^2$ to obtain

$$r^2 + R^2 - a^2 = a^2 + R^2 - 2aR\cos\theta + R^2 - a^2$$

$$= 2R(R - a\cos\theta).$$

Then

$$R - a\cos\theta = \frac{r^2 + R^2 - a^2}{2R} .$$

Substituting this expression into dF above, we obtain

$$dF = Gm\frac{2\pi at\rho\, dr}{Rr}\;\frac{r^2 + R^2 - a^2}{2rR}$$

$$= Gm\frac{\pi at\rho(r^2 + R^2 - a^2)\, dr}{R^2r^2}$$

$$= \frac{Gm\pi at\rho}{R^2}\left[1 + \frac{R^2 - a^2}{r^2}\right] dr ;$$

$$F = \frac{Gm\pi at\rho}{R^2}\int_{R-a}^{R+a}\left[1 + \frac{R^2 - a^2}{r^2}\right] dr$$

$$= \frac{Gm\pi at\rho}{R^2}\left[R + a - R + a - (R^2 - a^2)\left(\frac{1}{R+a} - \frac{1}{R-a}\right)\right]$$

$$= \frac{Gm\pi at\rho}{R^2}\left[2a - (R - a) + R + a\right]$$

$$= \frac{Gm(4\pi a^2 t\rho)}{R^2}$$

The mass M of the complete shell is $4\pi a^2 t\rho$, so we have

$$F = G\frac{mM}{R^2} .$$

The particle at O "sees" the spherical shell of mass M as a particle of mass M situated at the center of the shell. We can consider the spherical shell as composed of a number of concentric shells, each homogeneous--like an onion. If we add the attractions of all the shells, we get the result that a sphere composed of concentric shells, each with possibly different densities, behaves like a particle with a mass equal to that of the sphere, located at the center of the sphere.

Lecture 9: Comments on the Teaching of Calculus and its Applications to Students in the Life and Management Sciences

The following comments are based on a four-year study involving approximately 2400 students.

A student interested in the life or management sciences is generally very conscious of the "real world." He views his precollege mathematical training as part of a discipline which he accepted, but which had little, if anything, to do with his particular interests.

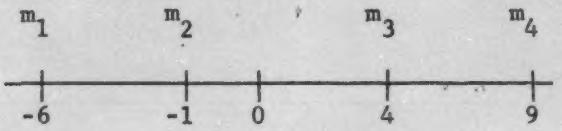
He considers the "applications" of mathematics presented in his secondary school training as either cute or fanciful. The reply to the question "What is mathematics?" generally alludes to "symbol playing" or "game playing." The result, of course, is that such a student is not motivated to continue a study of mathematics. If a teacher wishes to provide motivation for this student, it is of prime importance to assure him very early that mathematics does interact with the real world in a meaningful and significant way (in places other than the grocery store).

I have experimented with the following technique which, on the basis of student comments, I believe has been successful. On the first day of class I select a student and ask, "What are your interests?" The reply is generally a wide-ranging list which frequently includes--in addition to over-population, pollution, and money-making ideas--items such as photography, skin diving, drag-strip racing, etc. A teacher should be able to describe a situation in one of the areas of interest which is capable of mathematical treatment and a mathematical model should be presented. Relevant questions can now be obtained (or drawn) from the student. He is then assured that the richness of mathematics will aid him in getting information concerning answers to his questions. It is important to convince the student that such information will start coming "soon" and not one or two years hence. This means, of course, that some care must be exercised in the choice of the mathematical model presented. Many students are literally "fed-up" with being told, for example, that they should study quadratic equations because "some day in the future" they will make use of such a study.

It is an ideal situation if one can show the student (or, better yet, have the student show) that his previous mathematical training enables him to begin obtaining information immediately. However, it is important to consider questions which indicate a need for more powerful mathematical techniques.

Before proceeding, it is necessary to mention that the precalculus portion of our calculus course (for students in the life and management sciences) includes a one-half semester study of absolute value, graphing, and finite probability.

Example. Suppose a student indicates an interest in management decision-making (he seldom uses such language). One might then pose the following situation: There are four machines m_1, m_2, m_3, and m_4 in a certain factory which are located along an aisle (the x-axis) as illustrated in the figure below.



A new machine is to be located somewhere along this aisle. There will be interaction between the existing machines and the new machine. For example, items produced by the existing machines are brought to the new machine for further processing. Let us assume that the same number of items will be brought to the new machine from m_1 and from m_3. Six times as many items, however, will be brought to the new machine from m_2 as from m_1 and four times as many items will be brought to the new machine from m_4 as from m_1.

It is not difficult to draw a number of questions from students concerning this situation. One is invariably led to an assumption that there is a fixed cost of moving each item through one distance unit and to the question of where to locate the new machine in such a way that the total cost of the operation is a minimum. The mathematical model presented (after some discussion and simplification) is thus

$$c(x) = k\,d(x,-6) + 6k\,d(x,-1) + k\,d(x,4) + 4k\,d(x,9),$$

where k is the cost of transporting one item a unit distance, x represents the location of the new machine, and $d(x,-6)$, for example, represents the number of distance units between x and -6. Several interesting questions now arise:

i) Is the k-factor really necessary, or can we assume $k = 1$?

ii) What is $d(x,-6)$?

The notion of a function can also be discussed at this point.

Our problem, of course, is to determine a value of x which minimizes $c(x)$. (Problems such as this can also be valuable either now or later in showing a student that a solution to a problem is not necessarily unique.) Occasionally a student will suggest a method of solving this problem. If his suggestions are fruitful, one can then inject uncertainty into the situation. For example, suppose that the number of items moving from machine m_4 to the new machine varies with time--depending, say, upon the orders received for certain items. [An excellent discussion of the general deterministic machine-location problem appears in Noble, B. Applications of Undergraduate Mathematics in Engineering. New York, The Macmillan Company, 1967.]

The machine-location problem is merely one example; I am sure that most teachers could pose more interesting situations. It is important, however, that the student should not view the situation posed as, in reality, one which is "contrived," "artificial," or "unnatural." One reason I do make use of the machine-location problem is that I actually encountered a similar situation in a consulting experience.

The discussion thus far has concentrated on the initial exposure of a student (in the life or management sciences) to college mathematics. I have chosen to dwell on this subject because there is simply no reason that the illustrated process cannot, to a certain extent, continue. By no means am I suggesting that one devote an entire course, at this level, to building or studying mathematical models. Indeed, this would be wasteful. However, experience indicates that it is advantageous frequently to assure the student that the mathematics he is studying does, in a meaningful way, interact with the real world. The following is a partial outline of an attempt to make the student aware of such interaction:

1. Students in the life and management sciences need more than a superficial knowledge of exponential and logarithmic functions. The student soon becomes aware that

$$N'(t) = k\,N(t)$$

leads to

$$N(t) = A\,e^{kt}.$$

The classic growth problems (which appear, in some form, in every standard calculus text) can now be discussed in some detail. A number of limitations involved in the use of exponential models can be drawn from the students.

2. The student has been introduced to some elementary differential equations. Now, assume that the rate of increase in a certain population growth slows as the population approaches a "saturation level" S. One can give various reasons for this phenomenon, including lack of space, lack of food, etc. However, one can simply tell the student that certain populations have been observed to exhibit such behavior and that this phenomenon has been studied by biologists. (It is reassuring to the student when the teacher cites actual references.) The student is thus motivated to consider a differential equation of the form

$$N'(t) = k\,N(t)\left(1 - \frac{N(t)}{S}\right),$$

where k is a positive constant and $0 < N(t) < S$ for $t \geqq 0$. The behavior of a solution, the time at which the growth is increasing most rapidly, and the role of N(0) present interesting questions for the students.

3. At some stage the student should be exposed to the construction of a mathematical model built "from scratch." In Reference [4] we offer such a construction in considering the spread of an epidemic and in the birth-death processes. Although an understanding of such constructions requires some work, this is more "realistic" to the student than a simple offering of a differential equation such as (in a basic epidemic model)

$$P'(t) = k\,P(t)\,(1 - P(t)).$$

One should be very careful to point out the assumptions that are made in the construction of such models.

65

4. The discussions in Section 3 enable one to consider in a reasonable but elementary fashion the classic Volterra predator-prey model and the conclusions that are drawn concerning the use of certain types of insecticides. An elementary mathematical analysis is given in [4]. Confirmation of Volterra's principle is cited, for example, in [9].

A similar sequence of examples and models can be given for the management sciences.

Some sources of examples and models in the life and management sciences.

1. Allen, Roy G. Macro-Economic Theory. New York, St. Martin's Press, Inc., 1968.
2. Allen, Roy G. Mathematical Economics, 2nd ed. New York, St. Martin's Press, Inc., 1959.
3. Baumol, William J. Economic Theory and Operations Analysis, 2nd ed. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1965.
4. Baxter, Willard and Sloyer, Clifford. Calculus with Probability. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., to appear.
5. Brennan, Michael J., Jr. Preface to Econometrics, 2nd ed. Cincinnati, Ohio, South-Western Publishing Company, 1965.
6. Dean, Burton V., et al. Mathematics for Modern Management. New York, John Wiley and Sons, Inc., 1963.
7. Defares, J. G. and Sneddon, I. N. An Introduction to the Mathematics of Medicine and Biology. Chicago, Illinois, Year Book Medical Publishers, Inc., 1961.
8. Lotka, Alfred J. Elements of Mathematical Biology. New York, Dover Publications, 1957.
9. MacArthur, Robert H. and Connell, J. H. Biology of Populations. New York, John Wiley and Sons, Inc., 1966.
10. Pielou, E. C. An Introduction to Mathematical Ecology. New York, John Wiley and Sons, Inc., 1969.

11. Stern, Mark E. Mathematics for Management. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1963.

LINEAR ALGEBRA

Lecture 1: Introduction

Broadly speaking, we deal with two different types of "applications," namely applications in the other branches of mathematics and applications in the sciences. In the latter case there will be, in the background, a problem in the real world. Because most problems in the real world are too difficult to solve directly, one proceeds as follows:

"Real world"
↓
Abstraction and idealization
↓
Mathematical model
↓
Solution of mathematical problem
↓
Check relevance of solution to original problem

Moreover, the application must not be as remote from the students' knowledge as the topic which it illustrates.

It is useful to tabulate some broad areas of application. Many of these provide illustrations of a concrete nature involving matrices.

1. Applications to other branches of mathematics
    - (a) Linear equations
    - (b) Taylor series
    - (c) Systems of ordinary differential equations
2. Applications to computing
    - (a) Linear equations
    - (b) Eigenvalue problems

3. Statistics
    - (a) Least squares and regression
    - (b) Markov chains
4. Optimization
    - (a) Linear programming
    - (b) Nonlinear programming
    - (c) Theory of games
5. Physical sciences and engineering
    - (a) Dynamics of particles (e.g., satellite motion)
    - (b) Vibrating systems
    - (c) Structures
    - (d) Electrical networks
    - (e) Control theory
6. Nonphysical sciences
    - (a) Economics
    - (b) Social sciences (e.g., population shifts, factor analysis)
    - (c) Biological sciences

LINEAR ALGEBRA

Lecture 2: An Electrical Ladder Network

Mathematical Topic: Matrix multiplication

Comment: The matrix method provides a simple, natural, and systematic procedure for analysis of the network.

In mathematics classes we often motivate matrix multiplication by the following argument: Suppose that A is m × n, B is n × p, and

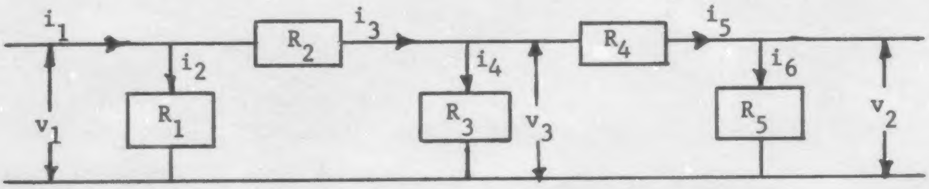
$$y = Bz \tag{1}$$

$$x = Ay, \tag{2}$$

where x, y, and z are $m \times 1$, $n \times 1$, and $p \times 1$, respectively. Then $x = A(Bz) = Cz$, say. We can find C either by eliminating y from (1) and (2) longhand or by the standard rule for matrix multiplication, i.e., $C = AB$. The expression for C is the same in both cases.

We discuss a physical example of an electrical ladder network in which multiplication of matrices arises naturally. Schematically a network may be represented by

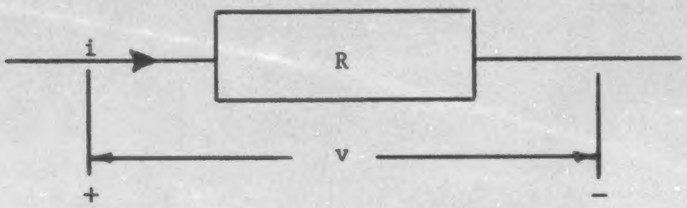
where i_1 and v_1 are the input current and voltage, respectively, and i_2 and v_2 are the output current and voltage, respectively. Within the "black box" may be other circuit elements, e.g., resistors. As a particular case we will examine the following network.



Here the i_k are currents, the v_k voltages, and the R_k resistances.

The physical principles which govern such a network are

1. Ohm's Law: for a single resistor as shown, $v = iR$

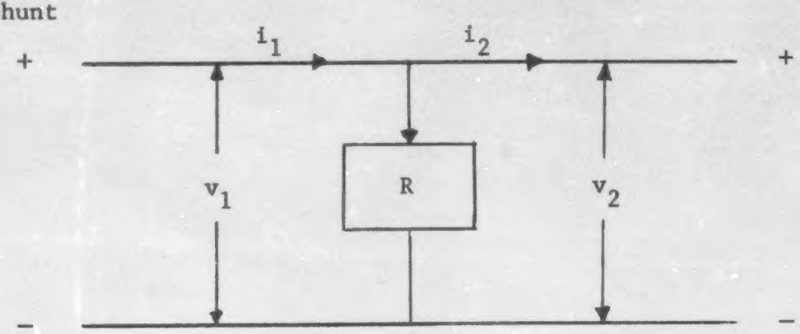


2. Kirchhoff's Law: the sum of the currents at a junction is zero

These laws may be applied directly to the network above only with some difficulty, resulting in a large number of equations. Compare that approach with the following matrix approach.

We first study two basic network blocks.

(a) Shunt



From Ohm's Law and Kirchhoff's Law we find

$$i = \frac{v_1}{R} = \frac{v_2}{R}$$

$$v_1 = v_2$$

$$i_1 = i_2 + i. \tag{3}$$

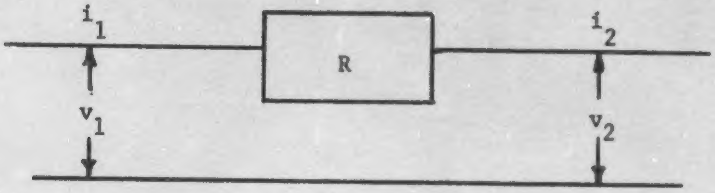
The equations (3) have the matrix expression

$$\begin{bmatrix} v_2 \\ i_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{R} & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ i_1 \end{bmatrix}. \tag{4}$$

With the notation $\underline{x}_2 = \begin{bmatrix} v_2 \\ i_2 \end{bmatrix}$, $\underline{x}_1 = \begin{bmatrix} v_1 \\ i_1 \end{bmatrix}$, and $P = \begin{bmatrix} 1 & 0 \\ -\frac{1}{R} & 1 \end{bmatrix}$, the shunt circuit may be summarized

$$\underline{x}_2 = P\underline{x}_1. \tag{5}$$

(b) Series



Here we have

$$i_1 = i_2$$

$$v_1 - v_2 = Ri_1 = Ri_2 \tag{6}$$

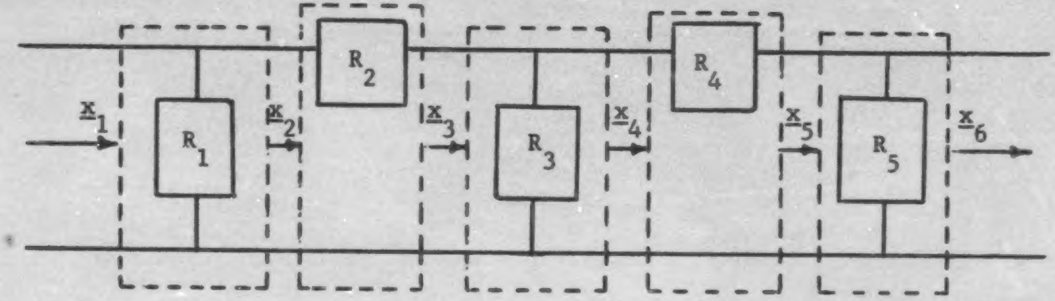
or

$$\begin{bmatrix} v_2 \\ i_2 \end{bmatrix} = \begin{bmatrix} 1 & -R \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ i_1 \end{bmatrix}. \tag{7}$$

With the notation $Q = \begin{bmatrix} 1 & -R \\ 0 & 1 \end{bmatrix}$, the series case may be summarized by the matrix equation

$$\underline{x}_2 = Q\underline{x}_1. \tag{8}$$

The original problem may now be viewed as blocks of these two basic types connected together.



Thus it is evident that

$$\begin{gathered}
\underline{x}_2 = A_1\underline{x}_1 \\
\underline{x}_3 = A_2\underline{x}_2 \\
\underline{x}_4 = A_3\underline{x}_3 \\
\underline{x}_5 = A_4\underline{x}_4 \\
\underline{x}_6 = A_5\underline{x}_5
\end{gathered} \tag{9}$$

and finally that

$$\underline{x}_6 = A_5A_4A_3A_2A_1\underline{x}_1, \tag{10}$$

where each A_k is either P or Q with the correct value for the resistance inserted.

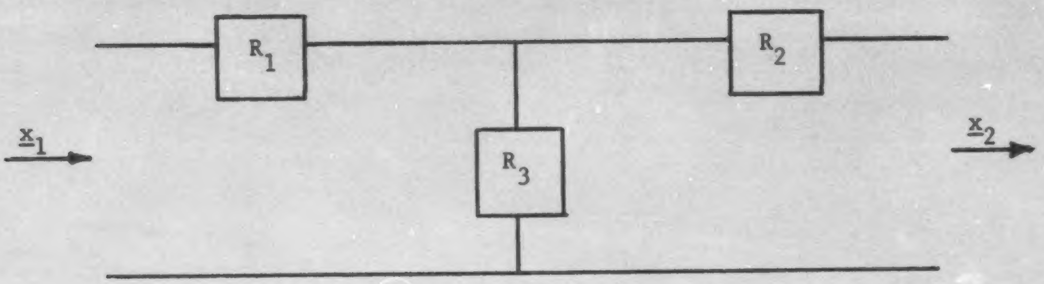
We have illustrated two main points.

a) Matrix multiplication occurs in a natural way in connection with electrical networks.

b) The matrix approach is simple and systematic. It proceeds by splitting the system into smaller subsystems, each of which has a simple matrix formulation. The analysis for the complete system is obtained by matrix multiplication.

Exercise: In each of the following, verify that $\underline{x}_2 = A\underline{x}_1$ for the given A.

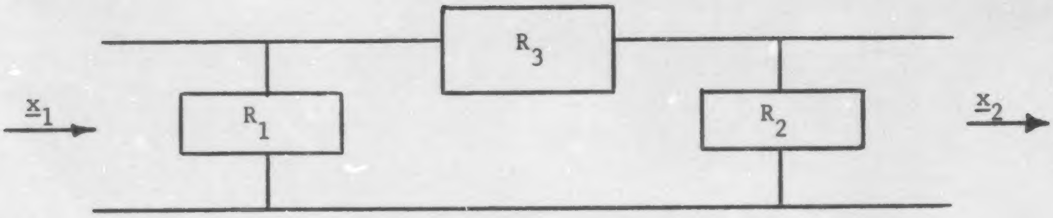
(a)



$$A = \begin{bmatrix} 1 + R_2Y_3 & -(R_1 + R_2 + R_1R_2Y_3) \\ - Y_3 & 1 + R_1Y_3 \end{bmatrix},$$

where $Y_3 = 1/R_3$

(b)



$$A = \begin{bmatrix} 1 + Y_1R_3 & - R_3 \\ -(Y_1 + Y_2 + Y_1Y_2R_3) & 1 + Y_2R_3 \end{bmatrix},$$

where $Y_1 = 1/R_1$ and $Y_2 = 1/R_2$

Notes and references

Many elementary textbooks in electronics contain treatments of ladder networks of the type discussed here, e.g., Nodelman and Smith [44]. The analysis can be carried much further when the network has a repetitive structure (e.g., when $R_{2i-1} = C_1$, $R_{2i} = C_2$, $i = 1, 2, \ldots$, where C_1 and C_2 are constants.)

Matrices can be applied to electrical networks that contain vacuum tubes, transistors, capacitors, inductors, etc., in quite general combinations (see, for example, Guillemin [28]).

Analogous applications of matrices occur in connection with mechanical structures. An elementary and instructive treatment of a simple example is given in Noble [43], pp. 27-39 (see also the references given there on p. 60).

LINEAR ALGEBRA

Lecture 3: Numerical Solutions of Linear Equations

Mathematical Topics: Linear equations, rank, row echelon form

Comment: In an abstract course one frequently omits many practical considerations which are relevant to the solution of a numerical problem. We shall therefore discuss some aspects of the numerical solution of simultaneous equations which are not ordinarily treated in a linear algebra course.

3.1 Mathematical theory

For the single equation in the unknown x

$$ax = b \tag{1}$$

one distinguishes three cases:

(a) $a \neq 0$: one solution

(b) $a = 0$, $b = 0$: infinitely many solutions

(c) $a = 0$, $b \neq 0$: no solution

In the $n \times m$ problem

$$A\underline{x} = \underline{b} \tag{2}$$

one develops analogous cases in terms of the rank of A or, equivalently, the dimension of the row space of A. The cases are:

(a) $r(A) = n$: one solution

(b) $r(A) = r[A,b] = k < n$: infinitely many solutions

(c) $r(A) < r[A,b]$: no solutions

In connection with case (b) one should note a pedagogic point. The general solution is

$$\underline{x} = \underline{x}_o + \sum_{i=1}^{n-k} d_i \underline{u}_i, \tag{3}$$

where the $\underline{u}_i$ are the basis of a certain vector space. This is a good illustration of the ideas connected with vector space and basis.

3.2 Practical methods

Digital computers usually solve linear equations by the method of successive, or Gaussian, elimination. As an example, consider

$$\begin{aligned} x_1 - x_2 + 2x_3 &= 6 \\ 2x_1 + x_2 - x_3 &= -1 \\ 5x_2 + x_3 &= -3. \end{aligned} \tag{4}$$

We can eliminate x_1 from the second and third equations by adding suitable multiples of the first equation to each of the other two. This gives

$$\begin{aligned} x_1 - x_2 + 2x_3 &= 6 \\ 3x_2 - 5x_3 &= -13 \\ 5x_2 + x_3 &= -3. \end{aligned} \tag{5}$$

The system (5), of course, has the same solution set as the original system (4). We now subtract 5/3 times the second equation from the third and get

$$\begin{aligned} x_1 - x_2 + 2x_3 &= 6 \\ 3x_2 - 5x_3 &= -13 \\ \frac{28}{3} x_3 &= \frac{56}{3}\ . \end{aligned} \tag{6}$$

We now find the unknowns by back substitution, the last equation giving x_3, then the second equation giving x_2, and finally x_1 resulting from the first.

Two points must be made in connection with the practical aspects of this procedure.

(a) The numbers 1, 3, $\frac{28}{3}$ which are the coefficients of x_1, x_2, x_3 in the first, second, and third equations, respectively, of (6) are called pivots. We would have obtained a different set of pivots if we had carried out the elimination in (4) in a different order. Note that the third equation of (4) cannot be used to eliminate x_1 from the first two equations. In practice it is advisable to adopt a partial pivoting method in which we inspect the coefficients of x_1 in the original set and choose the largest as a pivot. One proceeds similarly with x_2, x_3, ... In case (4) this means that the second

77

equation would be used to eliminate x_1. Similarly, in (5), 5 would be the pivot for x_2.

It is also the case that we can multiply any equation by a constant without changing the solution of the system. Similarly we could introduce new unknowns by setting $x_i = q_i y_i$, where the q_i are arbitrary nonzero constants. In terms of the matrix formulation $A\underline{x} = \underline{b}$, this means we can replace $A\underline{x} = \underline{b}$ by an equivalent system in which the rows and columns of $[A,b]$ are multiplied by arbitrary nonzero constants. In practice it is advisable to "normalize" or "scale" so that the largest element in each row and column is of order unity.

If we use partial pivoting, scaling can affect the choice of pivots. The reason for scaling in the way just described is that such scaling avoids numerical difficulties which occur if pivots are chosen incorrectly.

(b) Gaussian elimination is an efficient procedure. The number of operations involved in solving the single equations (2) is of order $\frac{1}{3} n^3$. To invert A (by Gaussian elimination) requires n^3 operations. The difference becomes quite striking for $n = 100$, say.

3.3 Ill-conditioning

A problem is said to be ill-conditioned if a "small" change in the coefficients causes a large change in the solution. Consider the system

$$\begin{aligned} x_1 + x_2 &= 2 \\ x_1 + 1.00001\, x_2 &= 2.00001 \end{aligned} \tag{7}$$

with solution

$$x_1 = 1,\ x_2 = 1. \tag{8}$$

The small change to the system

$$\begin{aligned} x_1 + x_2 &= 2 \\ x_1 + 1.00001\, x_2 &= 1.99990 \end{aligned} \tag{9}$$

changes the solutions to

$$x_1 = 12,\ x_2 = -10. \tag{10}$$

We may see how this happens by looking at the system more generally in the form

$$
\begin{aligned}
x_1 + x_2 &= 2 \\
x_1 + (1 + \varepsilon)x_2 &= 2 + \delta,
\end{aligned}
\qquad (11)
$$

where $\varepsilon \ll 1$. The solution is of the form

$$
x_2 = \frac{\delta}{\varepsilon}\ , \ x_1 = 2 - \frac{\delta}{\varepsilon}\ . \qquad (12)
$$

Now if $\varepsilon = 0$ and $\delta \neq 0$, there is no solution. If $\varepsilon = \delta = 0$, there is an infinity of solutions. If $\varepsilon \neq 0$ and $\delta \neq 0$, a change in δ is magnified by the factor $\frac{1}{\varepsilon}$. It is not generally possible to avoid small changes in the problem; these changes enter as rounding errors in the computer.

Another view of the trouble is provided by examining the determinant of the coefficients. We find

$$
\begin{vmatrix} 1 & 1 \\ 1 & 1 + \varepsilon \end{vmatrix} = \varepsilon, \qquad (13)
$$

which is nearly zero if ε is nearly zero. Thus a small determinant is a practical sign of ill-conditioning. Since the determinant is the product of the pivots, as may be seen in (6), small pivots are also a warning of possible trouble.

Small pivots may result from a rank deficiency in the matrix. For this circumstance some of the pivots should be zero but will probably not be exactly zero due to rounding error. The existence of some "reasonable" pivots together with some extremely small ones signals a possible rank deficiency. Consider the Hilbert matrix

$$
\begin{bmatrix}
1 & \frac{1}{2} & \cdots & \frac{1}{n} \\
\frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\
\cdot & \cdot & & \cdot \\
\cdot & \cdot & & \cdot \\
\cdot & \cdot & & \cdot \\
\frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n+1}
\end{bmatrix}.
$$

Here the pivots form a descending sequence, of the order $(1, 10^{-1}, 10^{-2}, \ldots)$.

3.4 Iterative improvement

Suppose we wish to solve $A\underline{x} = \underline{b}$ and $\hat{\underline{x}}$ is an approximate solution. The residual $\underline{r}$ is defined by

$$\underline{r} = \underline{b} - A\hat{\underline{x}}. \tag{14}$$

It seems reasonable to iterate in the following way: first subtract (14) from $A\underline{x} = \underline{b}$ to get

$$A(\underline{x} - \hat{\underline{x}}) = \underline{r}; \tag{15}$$

solve $A\underline{\delta} = \underline{r}$ to get $\hat{\underline{\delta}}$ and put

$$\begin{aligned} \underline{x} &= \hat{\underline{x}} + \underline{\delta} \\ &\cong \hat{\underline{x}} + \hat{\underline{\delta}} \end{aligned} \tag{16}$$

for an improved approximation. In fact, if $\hat{\underline{\delta}}$ is large the results are suspect and the improvements may well be meaningless. But the fact that $\hat{\underline{\delta}}$ is small provides a good indication that the original approximation was good.

The point is that $\underline{x} - \hat{\underline{x}} = A^{-1}\underline{r}$, so a small residual is meaningless if $|A^{-1}|$ is large. Thus a large value for $|A^{-1}|$ is another warning of ill-conditioning.

For the previous system (7), if

$$\hat{\underline{x}} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \tag{17}$$

then

$$\underline{r} = \begin{bmatrix} 0 \\ .00001 \end{bmatrix}. \tag{18}$$

But if

$$\hat{\underline{x}} = \begin{bmatrix} 1.01 \\ 1.01 \end{bmatrix}, \tag{19}$$

then

$$\underline{r} = \begin{bmatrix} -0.02 \\ -0.02 \end{bmatrix}. \tag{20}$$

Compare this with the exact solution (8).

We can see what happened if we write

$$\hat{\underline{x}} = \begin{bmatrix} 1 + \xi \\ 1 + \eta \end{bmatrix}; \tag{21}$$

this gives

$$\underline{r} = \begin{bmatrix} -(\xi + \eta) \\ -(\xi + 1.00001\eta) \end{bmatrix}. \tag{22}$$

Gaussian elimination automatically gives small r; this will be meaningful only for well-conditioned matrices.

3.5 Right and left inverses of square matrices

A related problem is the numerical realization of the following theorem: If C is the left inverse of the square matrix A, then C is the right inverse of A. In practice one considers C a good left inverse if I - CA is "small," and considers C a good right inverse if I - AC is small. A given matrix C need not be equally good as both left and right inverse. Consider

$$A = \begin{bmatrix} 1.00 & 1.00 \\ 1.00 & 0.99 \end{bmatrix}$$

$$C = \begin{bmatrix} -89 & 100 \\ 90 & -100 \end{bmatrix}$$

$$AC = \begin{bmatrix} 1.00 & 0.0 \\ 0.1 & 1.00 \end{bmatrix}$$

$$CA = \begin{bmatrix} 11 & 10 \\ -10 & -9 \end{bmatrix}.$$

Again, the reason for this numerical difficulty is not too hard to find.

Suppose

$$I - AC = \Delta, \quad \Delta \text{ small.} \tag{23}$$

Multiplying from the right by C^{-1} and from the left by C gives

$$I - CA = C\Delta C^{-1}. \tag{24}$$

Thus $C\Delta C^{-1}$ may not be small if C or C^{-1} has large elements.

3.6 Summary

We have emphasized the following points:

1. Gaussian elimination is an efficient numerical algorithm.
2. Some things to watch for in using a computer routine to implement the algorithm are
   (a) check the pivots to try to determine whether difficulties are rank deficiency or ill-conditioning
   (b) check the scaling
   (c) use iterative correction primarily as a cross-check.

3.7 Notes and references

Clearly, a detailed discussion of this topic is technical and belongs in a course in numerical analysis rather than a course in linear algebra. On the other hand, the student surely must be warned that, when using a computer, numerical examples do not necessarily work out according to the neat mathematical theory. The phenomenon of ill-conditioning is fundamental, yet this is not even mentioned in most textbooks on linear algebra.

An elementary reference is Forsythe and Moler [17]. A more advanced treatment is given in Noble [43].

LINEAR ALGEBRA

Lecture 4: Generalization from One Variable to n Variables; Norm

Mathematical Topics: Systems of algebraic equations, systems of differential equations, norms

Comment: We deal with some essentially mathematical applications of linear algebra in which we generalize from one variable to n variables. This procedure is affecting the curriculum in mathematics since it is elegant and natural to deal with multivariable calculus and systems of differential equations from this point of view. These ideas are also useful in connection with computing.

4.1 Straightforward iteration

Consider the problem of using the iteration method to find a number x such that

$$x = f(x). \tag{1}$$

From an initial guess $x^{(0)}$ successive approximations to x are produced according to

$$x^{(r+1)} = f(x^{(r)}). \tag{2}$$

If f obeys the proper Lipschitz conditions, the process will converge. We make this precise.

Theorem 4.1. Suppose that I is an interval containing a solution x of (1); $x^{(0)}$ is in I; for each positive integer r, $x^{(r)}$ is given by (2); $M < 1$; and for each a and b in I, $|f(a) - f(b)| \leq M|a - b|$. Then $\lim_{r\to\infty} x^{(r)} = x$.

Proof:
$$\begin{aligned} |x - x^{(r)}| &= |f(x) - f(x^{(r-1)})| \\ &\leq M\,|x - x^{(r-1)}| \\ &\leq M^r\,|x - x^{(0)}|, \end{aligned}$$

provided $x^{(r-1)}$ is in I. Since $x^{(0)}$ is in I and $M < 1$, it follows by induction that each $x^{(r)}$ is in I. Thus the preceding inequality is true for each r and, again since $M < 1$, the conclusion follows.

If one has a similar problem for several variables, i.e.,

$$
\begin{aligned}
x_1 &= f_1(x_1, \ldots, x_n) \\
x_2 &= f_2(x_1, \ldots, x_n) \\
&\;\;\vdots \\
x_n &= f_n(x_1, \ldots, x_n), \qquad (3)
\end{aligned}
$$

a matrix notation is obviously convenient. The equations (3) may be written

$$\underline{x} = \underline{f}(\underline{x}), \qquad (4)$$

where $\underline{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$, $\underline{f} = \begin{bmatrix} f_1 \\ \cdot \\ \cdot \\ \cdot \\ f_n \end{bmatrix}$.

The iteration becomes

$$\underline{x}^{(r+1)} = \underline{f}(\underline{x}^{(r)}). \qquad (5)$$

To generalize Theorem 4.1 to the multivariable case it will be necessary to introduce a quantity which corresponds to the absolute values which appear in that theorem. Such a quantity is the norm of a vector or matrix. Mathematically, a function which associates with a vector or matrix x a real number $\|x\|$ is a norm if

i) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$

ii) $\|kx\| = |k| \, \|x\|$ for every scalar k

iii) $\|x+y\| \leq \|x\| + \|y\|$.

Convenient norms for vectors and matrices are

$$\|x\| = \sup_i |x_i| \tag{6}$$

$$\|A\| = \sup_i \sum_j |a_{ij}|. \tag{7}$$

Note that (7) is equivalent to

$$\|A\| = \sup \|Ax\|, \ \|x\| = 1, \tag{8}$$

where (6) is used to evaluate the vector norms on the right hand side.

Properties i), ii), iii) of the norm may be readily verified. In addition it is immediate from (8) that

$$\|Ax\| \leq \|A\| \ \|x\|, \tag{9}$$

and it is also straightforward to show that, for matrices,

$$\|AB\| \leq \|A\| \ \|B\|. \tag{10}$$

There are many ways of defining norms; we have simply chosen one of these. A more general definition is to define, for any positive integer p,

$$\|x\|_p = \left(|x_1|^p + \ldots + |x_n|^p\right)^{1/p}. \tag{11}$$

The norm (6) is obtained by letting p tend to infinity in this formula, so the norm (6) is sometimes denoted by $\| \ \|_\infty$.

With the norm (6), for example, it is easy to generalize Theorem 4.1 to the multivariable case.

Theorem 4.2. Suppose that I is a rectangular region containing a solution $\underline{x}$ of (4); $\underline{x}^{(0)}$ is in I; for each positive integer r, $\underline{x}^{(r)}$ is given by (5); $M < 1$; and for each $\underline{a}$ and $\underline{b}$ in I, $\|\underline{f}(\underline{a}) - \underline{f}(\underline{b})\| \leq M \ \|\underline{a} - \underline{b}\|$. Then $\lim_{r \to \infty} \|x^{(r)} - x\| = 0$, i.e., $x^{(r)} \to x$ as $r \to \infty$.

The proof is analogous with that of Theorem 4.1. Note that, componentwise, $x_i^{(r)} \to x_i$.

To illustrate, we will apply these ideas to the solution of $A\underline{x} = \underline{b}$. To pose this problem we note that the solution $\underline{x}$ must obey

$$x_1 = -\frac{1}{a_{11}}(a_{12}x_2 + \dots + a_{1n}x_n) + b_1$$

$$\vdots$$

$$x_n = -\frac{1}{a_{nn}}(a_{n1}x_1 + \dots + a_{n,n-1}x_{n-1}) + b_n. \tag{12}$$

In matrix form this may be written

$$\underline{x} = F\underline{x} + \underline{\beta} = \underline{f}(\underline{x}), \tag{13}$$

to be solved by the iterations

$$\underline{x}^{(r+1)} = F\underline{x}^{(r)} + \underline{\beta}. \tag{14}$$

To check the hypothesis of Theorem 4.2 we observe that

$$\begin{aligned} \|\underline{f}(\underline{a}) - \underline{f}(\underline{b})\| &= \|F\underline{a} - F\underline{b}\| \\ &\leq \|F\| \, \|\underline{a} - b\|, \end{aligned} \tag{15}$$

the last step following from (9). Now

$$\|F\| = \sup_i \sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right|, \tag{16}$$

so convergence of (14) is assured if, for each i,

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|. \tag{17}$$

4.2 Newton's method

Newton's method of finding a root of $f(x) = 0$, given an approximate root $\hat{x}$, is to solve iteratively

$$x^{(r+1)} = x^{(r)} - \frac{f(x^{(r)})}{f'(x^{(r)})}, \tag{18}$$

with $x^{(0)} = \hat{x}$. To save space and avoid confusion due to details, we will give only a heuristic justification of (18). Let $x = \hat{x} + h$. By Taylor's theorem,

$$0 = f(\hat{x} + h) \approx f(\hat{x}) + hf'(\hat{x}), \tag{19}$$

so

$$h \approx -\frac{f(\hat{x})}{f'(\hat{x})}. \tag{20}$$

Thus

$$x \approx \hat{x} - \frac{f(\hat{x})}{f'(\hat{x})}, \tag{21}$$

to a first approximation. The procedure (18) continues in this manner.

The n-variable problem is

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0, \end{aligned} \tag{22}$$

or, in matrix notation,

$$\underline{f}(\underline{x}) = 0, \tag{23}$$

with an initial approximation $\hat{\underline{x}}$. With $\underline{x} = \hat{\underline{x}} + \underline{h}$ we have, using the n-variable Taylor series,

$$0 = f_i(\hat{\underline{x}} + \underline{h}) \approx f_i(\hat{\underline{x}}) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\hat{x})\, h_j. \tag{24}$$

If we define the matrix $f'(\underline{x})$ by

$$f'(\underline{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix},$$

then equation (24) is conveniently written

$$\underline{f}(\hat{\underline{x}}) + f'(\hat{\underline{x}})\, \underline{h} \approx 0. \tag{26}$$

Thus

$$\underline{h} \approx -[f'(\hat{\underline{x}})]^{-1}\, \underline{f}(\underline{x}) \tag{27}$$

and

$$\underline{x} \approx \hat{\underline{x}} - [f'(\hat{\underline{x}})]^{-1}\, \underline{f}(\hat{\underline{x}}); \tag{28}$$

this is parallel to (21), and an iterative scheme may be used which corresponds to (18).

Theorem 4.3. If $\|S\| < 1$, then $I - S$ is nonsingular.

Proof: If $I - S$ is singular there must exist a nonzero vector $\underline{x}$ such that

$$(I - S)\underline{x} = 0, \text{ or } \underline{x} = S\underline{x}. \tag{29}$$

Then, from (9),

$$\|X\| = \|S\underline{x}\| \leq \|S\| \, \|\underline{x}\|. \tag{30}$$

From this it follows that $\|S\| \geq 1$, contrary to our hypothesis. Other proofs follow (a) from the fact that the largest eigenvalue is less than or equal to $\|S\|$, and (b) from investigation of the convergence of the series $I + S + S^2 + \ldots$ to $(I - S)^{-1}$.

The computer frequently produces an "inverse" to a matrix which is really singular, due to the presence of round-off errors. The following is a useful test.

Theorem 4.4. Define $R = I - AC$ or $R = I - CA$. If $\|R\| < 1$, then:

(a) A and C are both nonsingular

(b) $\|A^{-1}\| \leq \dfrac{\|C\|}{1 - \|R\|}$

Proof: Let $R = I - AC$. Then $AC = I - R$, and hence $(AC)^{-1}$ exists by Theorem 4.3. This means that both A and C possess inverses. Consideration of the series for $(I - R)^{-1}$ gives the estimate $\|(I - R)^{-1}\| \leq (1 - \|R\|)^{-1}$, from which the second part follows.

The norm permits an investigation of the conditioning of the problem $A\underline{x} = \underline{b}$. Let $A(\underline{x} + \underline{\xi}) = \underline{b} + \underline{\delta}$. Then

$$A\underline{\xi} = \underline{\delta}$$

$$\underline{\xi} = A^{-1}\underline{\delta}$$

and

$$\|\underline{\xi}\| \leq \|A^{-1}\| \, \|\underline{\delta}\|. \tag{31}$$

Also,

$$\|A\| \, \|\underline{x}\| \geq \|\underline{b}\|, \tag{32}$$

so that

$$\frac{\|\underline{\xi}\|}{\|\underline{x}\|} \leq \|A\| \, \|A^{-1}\| \, \frac{\|\underline{\xi}\|}{\|\underline{b}\|}. \tag{33}$$

Thus the problem is well-conditioned (see Section 3.3) if $\|A\| \, \|A^{-1}\|$ is small. Experience suggests that if this condition number $\|A\| \, \|A^{-1}\|$ is large, the problem is ill-conditioned.

4.4 Systems of ordinary differential equations

The initial value problem for one function

$$x'(t) = a(t)\, x(t) + g(t)$$

$$x(t_o) = x_o \tag{34}$$

is readily generalized to a system of differential equations

$$x_1'(t) = \sum_j a_{1j}(t)\, x_j(t) + g_1(t)$$

$$\vdots$$

$$x_n'(t) = \sum_j a_{nj}(t)\, x_j(t) + g_n(t)$$

$$x_i(t_o) = x_{io}. \tag{35}$$

As throughout this section, the system of n equations takes on the form of a single equation when a matrix notation is used:

$$\underline{x}'(t) = \underline{A}(t)\, \underline{x}(t) + \underline{g}(t)$$

$$\underline{x}(t_o) = \underline{x}_o$$

This problem has a unique solution for some region, $0 \leq t \leq T$, if all the elements of A and $\underline{g}$ are continuous. When $\underline{g}(t)$ is identically zero, the solutions form a vector space.

A higher-order equation for one variable may be reduced to a system of first-order equations. Thus a single computer program for first-order systems of equations may replace many special subroutines for higher-order equations.

Notes and references

Forsythe and Moler [17] give an elementary introduction to the use of norms in numerical analysis in connection with linear algebraic equations. For a more advanced treatment, see Noble [43].

Systems of differential equations are treated in Brauer, et al. [10]. This text exemplifies a trend away from the study of special cases of differential equations towards a general approach in terms of systems.

The book of Crowell and Williamson [12] gives an example of the way in which material covered in an elementary linear algebra course can be applied to advanced calculus.

LINEAR ALGEBRA

Lecture 5: Applications of Systems of Differential Equations with Constant Coefficients

Mathematical Topics: Eigenvalues, eigenvectors, basis of a vector space, similarity transformations

Comment: This section consists of a number of more or less disconnected notes that supplement and extend the treatment of the mechanical vibrating system given in the last chapter.

5.1 The solution of $\underline{x}''(t) = A\underline{x}(t)$

Suppose we wish to solve

$$\underline{x}''(t) = A\underline{x}(t), \tag{1}$$

where A is an n × n matrix of constants. On the basis of certain fundamental ideas connected with eigenvalues and eigenvectors, it is easy to see the form which $\underline{x}$ must take. Assume that A has n linearly independent eigenvectors $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$, with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, i.e.,

$$A\underline{x}_k = \lambda_k \underline{x}_k, \quad k = 1, \ldots, n. \tag{2}$$

If we set

$$p = [\underline{x}_1, \ldots, \underline{x}_n], \tag{3}$$

then it follows that

$$\begin{aligned} AP &= [A\underline{x}_1, \ldots, A\underline{x}_n] \\ &= [\lambda_1\underline{x}_1, \ldots, \lambda_n\underline{x}_n] \\ &= P\Lambda, \end{aligned} \tag{4}$$

where

$$\Lambda = \text{diag}\ (\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{5}$$

Since the $\underline{x}_i$ are linearly independent, P^{-1} exists; thus

$$P^{-1}AP = \Lambda$$

and

$$P\Lambda P^{-1} = A. \tag{6}$$

This can be used to rewrite (1) as

$$\underline{x}'' = P\Lambda P^{-1}\underline{x}, \tag{7}$$

from which

$$P^{-1}\underline{x}'' = \Lambda P^{-1}x. \tag{8}$$

Upon setting

$$\underline{y} = P^{-1}\underline{x},$$

we obtain

$$\underline{y}'' = \Lambda\underline{y}. \tag{9}$$

Since Λ is diagonal, equation (9) represents the n independent equations

$$y_k'' = \lambda_k y_k \tag{10}$$

with solutions

$$y_k(t) = A_k e^{\sqrt{\lambda_k}t + \alpha_k}. \tag{11}$$

The solution to (1) is obtained as

$$\underline{x} = P\underline{y}. \tag{12}$$

This analysis will fail if there are not n linearly independent eigenvectors; the Jordan canonical form is needed to handle the general case. This fact may be used to motivate the study of the Jordan canonical form. The sort of difficulty which may arise is illustrated by the matrix

$$A = \begin{bmatrix} \alpha & 1 \\ 0 & \alpha \end{bmatrix}. \tag{13}$$

This has a repeated eigenvalue, α, which causes no difficulty, but it has only one eigenvector, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

5.2 Positivity of eigenvalues

In many physical problems it is important to determine whether the eigenvalues are all positive. It is sometimes possible to investigate this from the point of view of the positive-definiteness of the matrix. The matrix A is positive-definite if

$$\begin{aligned} &\underline{y}^T A\underline{y} \geq 0 \text{ for all } \underline{y}; \\ &\underline{y}^T A\underline{y} = 0 \text{ only if } \underline{y} = 0. \end{aligned} \tag{14}$$

Suppose that A is positive-definite and that

$$A\underline{x}_i = \lambda_i \underline{x}_i, \ \underline{x}_i \neq 0. \tag{15}$$

We have

$$\underline{x}_i^T A\underline{x}_i = \lambda_i \underline{x}_i^T x_i \tag{16}$$

and, since $\underline{x}_i^T \underline{x}_i$ is positive for every nonzero vector $\underline{x}_i$, it follows that λ_i is positive.

As an illustration of the computation, we examine the matrix

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}. \tag{17}$$

If $\underline{y} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$ is any nonzero real vector, then

$$\begin{aligned} \underline{y}^T A\underline{y} &= a(2a - b) + b(-a + 2b - c) + c(-b + c) \\ &= a^2 + (a - b)^2 + (a - c)^2 > 0. \end{aligned} \tag{18}$$

Thus A has positive eigenvalues.

5.3 Homogeneous systems of linear equations

When considering eigenvalues we need to use the result that a system of homogeneous equations $B\underline{x} = 0$ has either infinitely many solutions or only the trivial solution $\underline{x} = 0$. We note in passing that in a course in elementary linear algebra this result should be deduced when dealing with the inhomogeneous system $B\underline{x} = \underline{b}$. In this general case there are three possibilities:

(a) $r(B) = n$: one solution

(b) $r(B) = r[B,b] < n$: infinitely many solutions

(c) $r(B) < r[B,b]$: no solution

Depending on the way in which this theory has been developed, it should be a trivial matter to show for $\underline{b} = 0$ that case (c) cannot occur and that in case (a) the unique solution is $\underline{x} = 0$.

5.4 Use of the characteristic polynomial in practice

When introducing eigenvalues there is a tendency to imply that using the characteristic polynomial is a convenient method for computing eigenvalues. This tendency is reinforced by the fact that practical methods for computing eigenvalues efficiently tend to be rather sophisticated--too sophisticated to be treated in an elementary course.

In an elementary course one tends to teach the iterative method of finding the largest or smallest eigenvalue. It should be emphasized that usually we do this because the analysis is instructive, not because we would use the iterative method in practice.

In practice we would normally find eigenvalues by using a computer subroutine prepared by experts. The best practical numerical method for finding the eigenvalues of a symmetric matrix is generally agreed to be the method of Householder. The question of the best method to be used for a nonsymmetric matrix has not been settled.

5.5 The vibrating string with external forces

The problem of the motion of three beads of equal mass attached to an elastic string is discussed in detail in Noble [43]. When a force of frequency ω is applied to each mass, the equations of motion are

$$kx_1'' = -3x_1 + x_2 + F_1e^{i\omega t}$$

$$kx_2'' = x_1 - 2x_2 + x_3 + F_2e^{i\omega t}$$

$$kx_3'' = x_2 - 3x_3 + F_3e^{i\omega t}, \qquad (19)$$

where k is a constant depending upon the mass. We investigate solutions which are proportional to exp (iωt) by looking for solutions of the form

$$x_j(t) = X_je^{i\omega t}, \quad j = 1, 2, 3, \qquad (20)$$

where the X_j are (in general, complex) constants. When (20) is substituted into (19) the result is

$$(A - \lambda I)\underline{x} = \underline{f}, \qquad (21)$$

where

$$\lambda = k\omega^2$$

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\underline{f} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix}$$

$$A = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}. \qquad (22)$$

Here we have an inhomogeneous equation for $\underline{x}$ given in terms of the parameter λ, which is determined by the frequency of the impressed forces. The problem is to find how the solution $\underline{x}$ depends upon λ.

We consider the general situation, assuming again that A has three linearly independent eigenvectors $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$, corresponding to the eigenvalues λ_1, λ_2, λ_3, respectively. Thus $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$ are a basis of E_3 and we can write

$$\underline{x} = \sum_i \alpha_i \underline{x}_i, \tag{23}$$

for some set of scalars α_i, and similarly

$$\underline{f} = \sum_i f_i \underline{x}_i. \tag{24}$$

Substitution of (23) and (24) in (21) gives

$$\sum_i \alpha_i (\lambda_i - \lambda) \underline{x}_i = \sum_i f_i \underline{x}_i \tag{25}$$

and hence, since the $\underline{x}_i$ are linearly independent,

$$\alpha_i (\lambda_i - \lambda) = f_i, \quad i = 1, 2, 3. \tag{26}$$

When $\lambda = \lambda_i$, there is no solution if $f_i \neq 0$ and infinitely many solutions if $f_i = 0$. When $\lambda \neq \lambda_i$, (26) may be solved for α_i, and the explicit form of the solution is

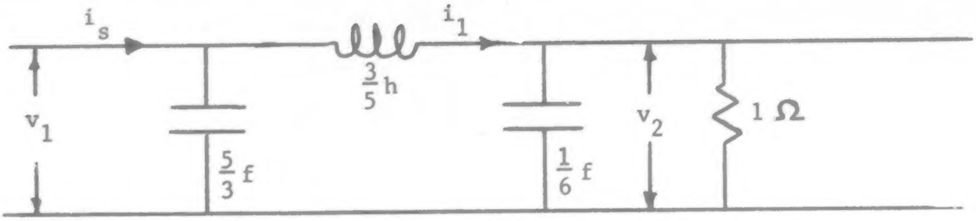
$$\underline{x}(\lambda) = \sum_i \left[\frac{f_i}{(\lambda_i - \lambda)} \right] \underline{x}_i. \tag{27}$$

Apart from the eigenvalues and eigenvectors of A, the only quantities needed to evaluate (27) are the coefficients f_i of the expansion of $\underline{f}$ in terms of the $\underline{x}_i$, given in (24). This equation may be solved as a linear system for the f_i. However, it is extremely convenient if the $\underline{x}_i$ have been chosen to be orthogonal. In this case f_i is readily computed as

$$f_i = \frac{\underline{x}_i^T \underline{f}}{\left(\underline{x}_i^T \underline{x}_i\right)^{\frac{1}{2}}}. \tag{28}$$

5.6 An L-C-R circuit

Consider the circuit



The problem is to find v_1, i_1, and v_2, given i_s. The circuit laws are

(a) capacitor: $i(t) = Cv'(t)$

(b) inductor: $v(t) = Li'(t)$

(c) resistor: $v(t) = Ri(t)$.

Application of these laws gives the equations

$$\begin{aligned} v_2 - v_1 &= \tfrac{3}{5} i_1{}' \\ \tfrac{1}{6} v_2{}' &= i_1 - v_2 \\ \tfrac{5}{3} v_1{}' &= -i_1 + i_s. \end{aligned} \tag{29}$$

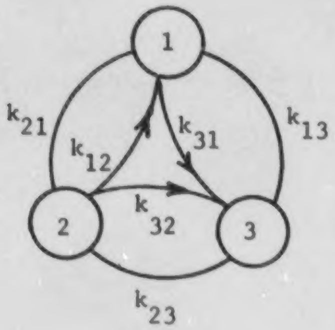
The equivalent matrix expression is

$$\begin{bmatrix} v_1{}' \\ i_1{}' \\ v_2{}' \end{bmatrix} = \begin{bmatrix} 0 & -\frac{3}{5} & 0 \\ \frac{5}{3} & 0 & -\frac{5}{3} \\ 0 & 6 & -6 \end{bmatrix} \begin{bmatrix} v_1 \\ i_1 \\ v_2 \end{bmatrix} + \begin{bmatrix} \frac{3}{5} \\ 0 \\ 0 \end{bmatrix} i_s. \tag{30}$$

(The Laplace method is an alternative approach, leading to the solution of polynomial equations.)

5.7 A chemical reaction with three components

A unimolecular chemical reaction may be described in terms of a set of constants k_{ji}, where k_{ji} is the rate constant for the reaction which converts species i to species j.



The amount of the i^{th} component present at time t, $a_i(t)$, is governed by the system

$$a_1' = -(k_{21} + k_{31})a_1 + k_{12}a_2 + k_{13}a_3$$

$$a_2' = k_{21}a_1 - (k_{12} + k_{32})a_2 + k_{23}a_3$$

$$a_3' = k_{31}a_1 + k_{32}a_2 - (k_{13} + k_{23})a_3,$$

or

$$\underline{a}'(t) = A\,\underline{a}(t). \tag{32}$$

Note that A is singular since the sum of each of its columns is zero.

The system (32) has a solution of the form

$$\underline{a}(t) = \underline{\alpha}_1 e^{\mu t} + \underline{\alpha}_2 e^{\nu t} + \underline{\alpha}_3 e^{\omega t}, \tag{33}$$

where μ, ν, ω are the eigenvalues of A. One of these, say ω, is zero since A is singular. The other two can be shown to be negative. Thus

$$\underline{a}(t) \to \underline{\alpha}_3 \quad \text{as} \quad t \to \infty.$$

It is frequently of interest to try to use experimental observations of the $a_i(t)$ together with equation (33) to find the rate constants. This is an ill-conditioned problem in practice; usually several μ, ν will fit the data about equally well. However, mathematics can be used to guide the experiment in such a way that the various exponentials are isolated, greatly reducing the attendant difficulties. (See section 5.9.)

5.8 Exponential of a matrix

The exponential of a matrix is crucial to the solution of linear systems

$$\underline{x}' = A\underline{x}, \tag{34}$$

as would be expected from analogy with the single-variable problem. To define the exponential or other transcendental functions of a matrix requires the investigation of the convergence of an infinite sequence of matrices. This investigation is conveniently carried out by means of the norm.

Given a sequence of matrices B_1, B_2, ..., if $\lim_{k\to\infty} \|B_k - B\| = 0$, then the sequence of (i,j) components of B_k converges to the (i,j) component of B. Thus we define $\lim_{k\to\infty} B_k = B$ to mean that $\lim_{k\to\infty} \|B_k - B\| = 0$.

Let P be an arbitrary square matrix. To study the convergence of the series

$$I + P + \frac{P^2}{2!} + \dots \tag{35}$$

we examine the sequence $S_1, S_2, \dots$ of partial sums

$$S_k = I + P + \dots + \frac{P^k}{k!} \, . \tag{36}$$

We note that, if $m > n$,

$$S_m - S_n = \frac{P^{n+1}}{(n+1)!} + \dots + \frac{P^m}{m!} \tag{37}$$

and

$$\|S_m - S_n\| \le \frac{\|P\|^{n+1}}{(n+1)!} + \dots + \frac{\|P\|^m}{m!} \, . \tag{38}$$

Thus convergence follows from the Cauchy criterion, and the proof mimics the scalar case.

If P can be diagonalized by a similarity transformation B, i.e.,

$$\Lambda = BPB^{-1} \tag{39}$$

where

$$\Lambda = \mathrm{diag}\ (\lambda_1, \dots, \lambda_n), \tag{40}$$

then e^P may be easily computed as

$$e^P = B^{-1}\ \mathrm{diag}\ (e^{\lambda_1}, \dots, e^{\lambda_n})\ B. \tag{41}$$

In the latter form it is easy to see that $\underline{x}' = A\underline{x}$ has solution

$$\underline{x}(t) = e^{At}\underline{x}_o. \tag{42}$$

One computes

$$\begin{aligned} \underline{x}'(t) &= \left(\frac{d}{dt}\, e^{At}\right)\underline{x}_o \\ &= \left(\frac{d}{dt}\, B^{-1}\ \mathrm{diag}\ (e^{\lambda_1 t}, \dots, e^{\lambda_n t})\ B\right)\underline{x}_o \\ &= \left(B^{-1}\, \Lambda\ \mathrm{diag}\ (e^{\lambda_1 t}, \dots, e^{\lambda_n t})\ B\right)\underline{x}_o \\ &= (B^{-1}\, \Lambda\, B)\left(B^{-1}\ \mathrm{diag}\ (e^{\lambda_1 t}, \dots, e^{\lambda_n t})\ B\right)\underline{x}_o \\ &= A\, e^{At}\underline{x}_o \\ &= A\, \underline{x}(t). \end{aligned} \tag{43}$$

5.9 Notes and references

Considerable additional detail on most of the topics in this chapter can be found. For section 5.1 see Brauer, et al. [10] and Bellman [5]. For 5.5, 5.6, 5.7 a complete discussion may be found in Noble [42]. Bellman [5] may be consulted relative to 5.8.

LINEAR ALGEBRA

Lecture 6: Diagonalization and Quadratic Forms

Mathematical topics: Diagonalization, quadratic forms

Comment: Applications of quadratic forms occur in several physical contexts. These can be used to provide examples involving eigenvalues and eigenvectors that are different from those which arise in connection with systems of differential equations. Due to lack of time, there is a tendency to omit quadratic forms from elementary linear algebra courses, but the topic is important in applications.

6.1 Introduction

A quadratic form is an expression $\underline{x}^T A \underline{x}$, where A is a real symmetric matrix:

$$\underline{x}^T A \underline{x} = \sum_{i,j} a_{ij} x_i x_j . \tag{1}$$

Quadratic forms occur in many contexts, including the following:

(1) Elementary geometry (Study of a quadratic form makes possible the classification of conic sections.)

(2) Taylor series (When f(x,y) has the series

$$\begin{aligned} f(x,y) = f(x_o,y_o) &+ (x - x_o)\, f_x(x_o,y_o) + (y - y_o)\, f_y(x_o,y_o) \\ &+ \tfrac{1}{2}(x - x_o)^2\, f_{xx}(x_o,y_o) + (x - x_o)\,(y - y_o)\, f_{xy}(x_o,y_o) \\ &+ \tfrac{1}{2}(y - y_o)^2\, f_{yy}(x_o,y_o) + \dots \quad , \end{aligned} \tag{2}$$

the second-order terms comprise a quadratic form which is studied to determine maxima and minima.)

(3) Optimization (closely connected with (2))

(4) Small vibrations

(5) Moments of inertia

(6) Statistics (regression, factor analysis, analysis of variance)

Several of these topics will be examined in greater detail.

101

6.2 Elementary Geometry

The type of conic section represented by an equation of the form

$$ax^2 + 2kxy + by^2 = c \tag{3}$$

is not obvious from an examination of the coefficients. The situation is much clearer, however, if the equation for the curve is expressed in terms of coordinates X, Y with respect to which there are no cross-terms.

For this two-variable case, X and Y can be found by longhand methods. If we set

$$\begin{aligned} x &= X \cos \theta - Y \sin \theta \\ y &= X \sin \theta - Y \cos \theta \ , \end{aligned} \tag{4}$$

then equation (3) becomes

$$\alpha X^2 + \beta Y^2 = c \ , \tag{5}$$

where $\tan 2\theta = \frac{2k}{a-b}$. The values of α and β determine the type of curve.

These results may be expressed from the matrix point of view with the introduction of the following:

$$\underline{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \ A = \begin{bmatrix} a & k \\ k & b \end{bmatrix}, \ S = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$\underline{X} = \begin{bmatrix} X \\ Y \end{bmatrix}, \ D = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \tag{6}$$

Equation (3) may be written

$$\underline{x}^T A \underline{x} = c, \tag{7}$$

equation (4) is

$$\underline{x} = S\underline{X}, \tag{8}$$

and equation (5) becomes

$$\underline{X}^T D \underline{X} = c. \tag{9}$$

Substituting (8) in (7) gives

$$\underline{X}^T S^T A S \underline{X} = c. \tag{10}$$

Thus a solution to the problem is obtained if one finds an orthogonal matrix S such that $S^T AS$ is diagonal.

The matrix formulation readily generalizes to n dimensions and provides a motivation for the matrix study of quadratic forms. The eigenvalues of A determine the form of conic section of (1). The eigenvectors give the matrix S and hence determine the axes with respect to which the cross-terms vanish.

6.3 Review of mathematical theory

Let A be a real symmetric n × n matrix. We know from general theory that A has n linearly independent, mutually orthogonal eigenvectors $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. We introduce $S = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]$.

Theorem 6.1. $S^TS = I$ (hence $S^T = S^{-1}$) if the $\underline{x}_i$ are chosen to be orthonormal.

Proof:

$$S^TS = \begin{bmatrix} \underline{x}_1{}^T \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_n{}^T \end{bmatrix} [\underline{x}_1, \ldots, \underline{x}_n]$$

$$= \begin{bmatrix} \underline{x}_1{}^T\underline{x}_1 & \underline{x}_1{}^T\underline{x}_2 & \cdots & \underline{x}_1{}^T\underline{x}_n \\ \underline{x}_2{}^T\underline{x}_1 & \underline{x}_2{}^T\underline{x}_2 & \cdots & \underline{x}_2{}^T\underline{x}_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \underline{x}_n{}^T\underline{x}_1 & \underline{x}_n{}^T\underline{x}_2 & \cdots & \underline{x}_n{}^T\underline{x}_n \end{bmatrix} = I.$$

Theorem 6.2. Let Λ be a diagonal matrix whose i^{th} diagonal element is λ_i, the eigenvalue corresponding to $\underline{x}_i$. Then $AS = S\Lambda$.

Proof:

$$\begin{aligned} AS &= [A\underline{x}_1, \ldots, A\underline{x}_n] \\ &= [\lambda_1\underline{x}_1, \ldots, \lambda_n\underline{x}_n] \\ &= S\Lambda \end{aligned}$$

Theorem 6.3. If $\underline{x} = S\underline{y}$, then $\underline{x}^TA\underline{x} = \underline{y}^T\Lambda\underline{y} = \lambda_1y_1{}^2 + \ldots + \lambda_ny_n{}^2$.

103

Proof:

$$\underline{x}^T A \underline{x} = \underline{y}^T S^T A S \underline{y}$$
$$= \underline{y}^T \Lambda \underline{y}$$

6.4 An example from chemical engineering

A chemical reaction involves the temperature T, the concentration c of one of the reactants, and the time t of the reaction. The yield η of the reaction is to be investigated as T, c, and t are varied within certain limits. Preliminary experiments indicate that we are interested in values in the ranges $T = 167 \pm 5^\circ$ C, $c = 27.5 \pm 2.5\%$, $t = 6.5 \pm 1.5$ hr. We introduce the standardized variables $x_1 = \frac{T - 167}{5}$, $x_2 = \frac{c - 27.5}{2.5}$, $x_3 = \frac{t - 6.5}{1.5}$ which have range of interest between -1 and 1.

Experimental observation gives

$$\eta = 57.79 + 1.78\, x_1 + 0.50\sqrt{3}\, x_2 + 1.08\sqrt{3}\, x_3 - 1.4\, x_1^2 - 1.4\, x_3^2 - 1.4\sqrt{3}\, x_1 x_2 - 1.2\sqrt{3}\, x_1 x_3 - 1.2\, x_2 x_3. \quad (11)$$

This is of the form

$$\eta = \underline{x}^T A \underline{x} - 2\underline{x}^T \underline{b} + c. \quad (12)$$

We remove the linear terms by setting $\underline{x} = \underline{z} + \underline{z}_o$. This gives

$$\eta = (\underline{z} + \underline{z}_o) A (\underline{z} + \underline{z}_o) - 2(\underline{z} + \underline{z}_o)^T \underline{b} + c \quad (13)$$
$$= \underline{z}^T A \underline{z} + 2\underline{z}^T (A\underline{z}_o - \underline{b})$$
$$+ c - 2\underline{z}_o^T \underline{b} + \underline{z}_o^T A \underline{z}_o,$$

and the linear terms will vanish if $\underline{z}_o$ is chosen to satisfy

$$A\underline{z}_o = \underline{b}. \quad (14)$$

Equation (13) then becomes

$$\eta = \underline{z}^T A \underline{z} + (c - \underline{z}_o^T A \underline{z}_o).$$

In the case of (11), this procedure results in

$$\eta = 58.44 - 1.4\, z_1^2 - 1.4\, z_3^2 - 1.4\sqrt{3}\, z_1 z_2$$

$$- 1.2\sqrt{3}\, z_1 z_3 - 1.2\, z_2 z_3$$

$$= 58.44 + \underline{z}^T A \underline{z}, \tag{15}$$

where

$$A = \begin{bmatrix} -1.4 & -0.7\sqrt{3} & -0.6\sqrt{3} \\ -0.7\sqrt{3} & 0 & -0.6 \\ -0.6\sqrt{3} & -0.6 & -1.4 \end{bmatrix} . \tag{16}$$

The eigenvalues of A are found to be $\lambda_1 = -3$, $\lambda_2 = -0.5$, $\lambda_3 = 0.7$, and

$$S = \begin{bmatrix} 0.4\sqrt{3} & 0.3\sqrt{3} & 0.5 \\ 0.4 & 0.3 & -0.5\sqrt{3} \\ 0.6 & -0.8 & 0 \end{bmatrix} .$$

Applying the diagonalization procedure to (16), we set $\underline{z} = S\underline{y}$ and find

$$\eta = 58.44 - 3y_1^2 - 0.5\, y_2^2 + 0.7\, y_3^2. \tag{17}$$

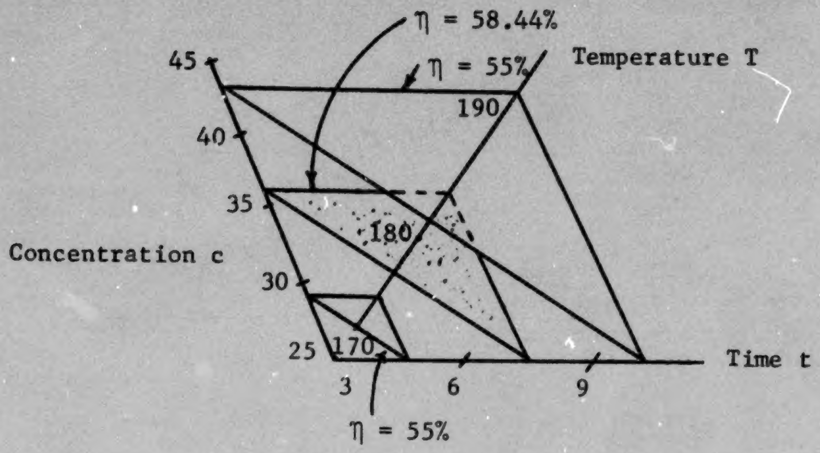
It can be checked that, to a crude approximation, we can neglect the last two terms in (17) (recall that the original variables were between 1 and -1) and consider

$$\eta \cong 58.44 - 3y_1^2. \tag{18}$$

The maximum of η is therefore given by $y_1 = 0$. Expressing this in terms of T, c, and t gives the condition

$$\sqrt{3}(T - 167) + 2(c - 27.5) + 5(t - 6.5) = 3.25\sqrt{3}. \tag{19}$$

This is a plane in T,c,t-space. The statement that the maximum η is given by $y_1 = 0$ is equivalent to the statement that the maximum is given by any point on the plane (19). This is illustrated in the figure by the shaded plane.



We also indicate two planes for which $\eta = 55\%$, slightly less than the maximum value.

The conclusion of the analysis is that, in the region of interest, the level surfaces are approximately planes.

6.5 Rayleigh's Principle

If A is a real symmetric n X n matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, then

$$\max \frac{\underline{x}^T A \underline{x}}{\underline{x}^T \underline{x}} = \lambda_n$$

and

$$\min \frac{\underline{x}^T A \underline{x}}{\underline{x}^T \underline{x}} = \lambda_1, \tag{20}$$

where the max and min are taken over all $\underline{x}$, or over all $\underline{x}$ such that $\underline{x}\underline{x}^T = 1$.

Proof: There are n linearly independent orthonormalized eigenvectors $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$. Since these vectors span the space, there are scalars α_i such that $\underline{x} = \sum_{i=1}^{n} \alpha_i \underline{x}_i$. Then

$$\frac{\underline{x}^T A \underline{x}}{\underline{x}^T \underline{x}} = \frac{\sum_{i=1}^{n} \lambda_i \alpha_i^2}{\sum_i \alpha_i^2}$$

and

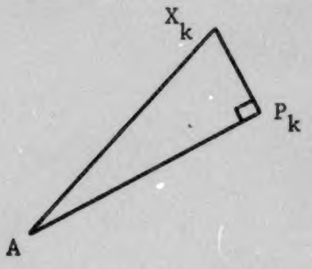
$$\frac{\underline{x}^T A \underline{x}}{\underline{x}^T \underline{x}} - \lambda_1 = \frac{\sum_{i=2}^{n} (\lambda_i - \lambda_1)\, \alpha_i^2}{\sum_i \alpha_i^2} \geq 0 ,$$

since λ_1 is the smallest eigenvalue. This expression is clearly minimized by $\underline{x}_1$. The other statement is proved similarly.

6.6 An application in statistics

Consider the following problem, which arises in several contexts in statistics, including factor analysis. We have points $X_1, X_2, \ldots, X_n$ in m-dimensional space, with coordinate $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$, respectively. We seek a straight line $\overline{CD}$ such that the sum of the squares of the distances between the points and the line is minimized.

Let the equation of $\overline{CD}$ be $\underline{x}(t) = t\underline{b} + \underline{a}$, where $\underline{b}^T\underline{b} = 1$. Let A be the fixed point on $\overline{CD}$ with coordinates $\underline{a}$ and let P_k be the foot of the perpendicular from X_k to $\overline{CD}$.



The square of the distance from A to X_k is $(\underline{\xi}_k - \underline{a})^T (\underline{\xi}_k - \underline{a})$, and the distance from A to P_k is $\underline{b}^T (\underline{\xi}_k - \underline{a})$. Therefore the square of the distance from X_k to the line $\overline{CD}$ is

$$(\underline{\xi}_k - \underline{a})^T (\underline{\xi}_k - \underline{a}) - [\underline{b}^T(\underline{\xi}_k - \underline{a})]^2. \qquad (23)$$

The sum of the squares of these distances, S^2, is given by

$$S^2 = \sum_{k=1}^{n} \Big((\underline{\xi}_k - a)^T(\underline{\xi}_k - \underline{a}) - [\underline{b}^T(\underline{\xi}_k - \underline{a})]^2 \Big). \qquad (24)$$

We choose $\underline{a}$ to minimize S^2, i.e., to make $\frac{\partial(S^2)}{\partial a_i} = 0$, $i = 1, \ldots, n$. After some calculation we find that

$$\underline{a} = \frac{1}{n} \sum_{k=1}^{n} \underline{\xi}_k + c\underline{b}, \tag{25}$$

where c is an arbitrary constant. It can be verified that the value of S^2 is independent of c. This corresponds geometrically to the fact that if $\underline{z}$ is on the line, then so is $\underline{z} + c\underline{b}$, since $\underline{b}$ specifies the direction of the line. For simplicity we choose c = 0, so

$$\underline{a} = \frac{1}{n} \sum_{k=1}^{n} \underline{\xi}_k. \tag{26}$$

We introduce $\hat{\underline{\xi}}_k = \underline{\xi}_k - \underline{a}$ and obtain

$$\begin{aligned} S^2 &= \sum_{k=1}^{n} \left(\hat{\underline{\xi}}_k^T \hat{\underline{\xi}}_k - (\underline{b}^T \hat{\underline{\xi}}_k)^2 \right) \\ &= \sum_{k=1}^{n} \left(\hat{\underline{\xi}}_k^T \hat{\underline{\xi}}_k - \underline{b}^T \hat{\underline{\xi}}_k \hat{\underline{\xi}}_k^T \underline{b} \right) \\ &= \sum_{k=1}^{n} \left(\hat{\underline{\xi}}_k^T \hat{\underline{\xi}}_k - \underline{b}^T A \underline{b} \right), \end{aligned} \tag{27}$$

$$\text{where } A = \sum_{k=1}^{n} \hat{\underline{\xi}}_k^T \hat{\underline{\xi}}_k^T. \tag{28}$$

Thus, from (27), to minimize S^2 is to maximize $\underline{b}^T A \underline{b}$ subject to $\underline{b}^T \underline{b} = 1$, when A is given by (28). This is precisely the problem in equations (20), considered in the last section.

6.7 Notes and references

The geometrical example in 6.2 is treated in detail from this point of view in Noble [43], pp. 378-385. A detailed account of 6.4 is found in the chapter by S. E. R. Box in Davies [14]. A brief treatment is given in Noble [43], pp. 388-390, and another treatment including an elementary discussion of the method of steepest descents is given in Noble [42], Chapter 3. The problem in 7.6 is treated, although in longhand fashion, in Hyvarinen [32], pp. 82-87.

LINEAR ALGEBRA

Lecture 7: Linear Programming

Mathematical Topics: Elementary row operations on matrices, the Gauss-Jordan method, elementary n-dimensional geometry, elementary probability

Comment: Linear programming deals with the maximization of a linear function subject to linear inequalities. It has many applications in industry and government. The way in which certain games of strategy can be formulated in terms of linear programming is also briefly described.

7.1 A typical linear programming problem

A factory makes two products, each of which utilizes three machines for its manufacture. We tabulate the hours needed on each machine for the two products and the total time available on each machine.

Machine	Product 1	Product 2	Hrs. Avail.
1	2	1	70
2	1	1	40
3	1	3	90

The profit per unit is \$40 for product 1 and \$60 for product 2. It is required to find the number of units x_1 and x_2 of products 1 and 2, respectively, which will maximize the profit M.

The profit is given by

$$M = 40x_1 + 60x_2 \tag{1}$$

subject to the conditions

$$\begin{aligned} &x_1 \geqq 0,\ x_2 \geqq 0 \\ &2x_1 + x_2 \leqq 70 \\ &x_1 + x_2 \leqq 40 \\ &x_1 + 3x_2 \leqq 90\,. \end{aligned} \tag{2}$$

This has a matrix expression of the form

$$M = \underline{c}^T\underline{x}$$

$$\underline{x} \geqq 0$$

$$A\underline{x} \leqq \underline{b} \, , \tag{3}$$

where, for matrices, $P \geqq Q$ means that each element of P is greater than or equal to the corresponding element of Q.

The maximization of M may be approached either algebraically or geometrically. We shall first examine the algebraic method.

7.2 An algebraic approach: the simplex method

The simplex method, discovered by Dantzig in 1947, is an instructive application of elementary row operations on matrices. It is important in practice because it is easy to computerize. We will demonstrate this technique by solving the problem of 7.1.

The first step is to convert $A\underline{x} \leqq \underline{b}$ into an equality $B\underline{z} = \underline{b}$ by the introduction of slack variables. In detail, we set

$$\begin{aligned} 2x_1 + x_2 + x_3 &= 70 \\ x_1 + x_2 + x_4 &= 40 \\ x_1 + 3x_2 + x_5 &= 90 \, . \end{aligned} \tag{4}$$

Our plan is to find a particular solution, which we will then improve to make M a maximum.

We introduce some convenient terminology. When B is $m \times n$, a solution with $n - m$ of the variables equal to zero is called a basic solution. A solution of $B\underline{z} = \underline{b}$ that satisfies $\underline{z} \geqq 0$ is called a feasible solution. A feasible solution that maximizes M is called an optimal feasible solution. A fundamental theorem, proved in [43], is

Theorem 7.1. If an optimal feasible solution exists, then a basic optimal feasible solution exists.

This theorem is important because it implies that optimal solutions are found among the basic solutions.

An obvious feasible solution of (4) is $x_1 = 0$, $x_2 = 0$, $x_3 = 70$, $x_4 = 40$, $x_5 = 90$. This gives $M = 40x_1 + 60x_2 = 0$. We now vary x_1 and x_2 and try to obtain a solution which maximizes M. We first vary x_2, since it has the larger coefficient, leaving x_1 at zero. Since $\underline{z} \geqq 0$ we are constrained by

$$\begin{aligned} x_3 &= 70 - x_2 \geqq 0 \\ x_4 &= 40 - x_2 \geqq 0 \\ x_5 &= 90 - 3x_2 \geqq 0 \end{aligned} \tag{5}$$

The largest x_2 consistent with all equations (5) is $x_2 = 30$. With this value of x_2, we have $x_1 = 0$, $x_2 = 30$, $x_3 = 40$, $x_4 = 10$, and $M = 1800$.

We were able to maximize x_2 readily because (1) M was expressed in terms of the zero variables and (2) each equation contained just one nonzero (or basic) variable. The second condition can be achieved for our current set of basic variables x_2, x_3, x_4 by applying elementary row operations to (4). We obtain the equivalent system

$$\begin{aligned} \tfrac{5}{3}x_1 \qquad\qquad + x_3 \qquad\qquad - \tfrac{1}{5}x_5 &= 40 \\ \tfrac{2}{3}x_1 \qquad\qquad\qquad\qquad + x_4 - \tfrac{1}{3}x_5 &= 10 \\ \tfrac{1}{3}x_1 + x_2 \qquad\qquad\qquad\qquad + \tfrac{1}{3}x_5 &= 30\ . \end{aligned} \tag{6}$$

M may now be written in terms of x_1 and x_5 as

$$\begin{aligned} M &= 40x_1 + 60(30 - \tfrac{1}{3}x_1 - \tfrac{1}{3}x_5) \\ &= 1800 + 20x_1 - 20x_5. \end{aligned} \tag{7}$$

To maximize M we should increase x_1, but leave x_5 at zero. From (6) we have the constraints

$$\begin{aligned} x_3 &= 40 - \tfrac{5}{3}x_1 \geqq 0 \\ x_4 &= 10 - \tfrac{2}{3}x_1 \geqq 0 \\ x_2 &= 30 - \tfrac{1}{3}x_1 \geqq 0. \end{aligned} \tag{8}$$

The maximum compatible value of x_1 is 15, which leads to $x_1 = 15$, $x_2 = 25$, $x_3 = 15$, $x_4 = 0$, $x_5 = 0$, and $M = 2100$.

If M is now expressed in terms of the zero variables x_4 and x_5, we find $M = 2100 - 30x_4 - 10x_5$. Thus M cannot be further increased and we are through, having found $x_1 = 15$, $x_2 = 25$, $M = 2100$.

The general case starts with the problem of maximizing

$$M = \underline{c}^T\underline{x} \tag{9}$$

subject to

$$A\underline{x} \leqq \underline{b} \text{ and } \underline{x} \geqq 0, \tag{10}$$

where A is $m \times n$, $\underline{x}$ is $n \times 1$, and $\underline{b}$ is $m \times 1$. We proceed as follows:

(1) Introduce slack variables $\underline{y} = [x_{n+1}, \ldots, x_m]^T$ to convert $A\underline{x} \leqq \underline{b}$ to

$$B\underline{z} = \underline{b}, \ \underline{z} = \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix}. \tag{11}$$

The problem is now to maximize

$$M = \underline{d}^T\underline{z}, \ d = \begin{bmatrix} \underline{c} \\ 0 \end{bmatrix} \tag{12}$$

subject to (11) and $\underline{z} \geqq 0$. We arrange B, $\underline{b}$, and $\underline{d}$ in the following array

$$P = \begin{bmatrix} B & \underline{b} \\ -\underline{d}^T & 0 \end{bmatrix}. \tag{13}$$

(2) Find a basic feasible solution of (11). In the preceding example this was done by inspection; methods are available for more difficult cases [43].

(3) Locate the largest negative number in the last row of P. The column in which it lies, say k, identifies x_k as the next variable to be introduced into the basis.

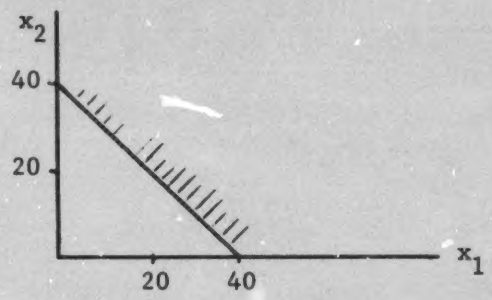
(4) Divide each element in the last column of P by the number in column k of the same row. Locate the divisor that yields the smallest quotient; this number will be used as the pivot in step (5).

(5) Pivot, using the Gauss-Jordan procedure, with the number located in step (4).

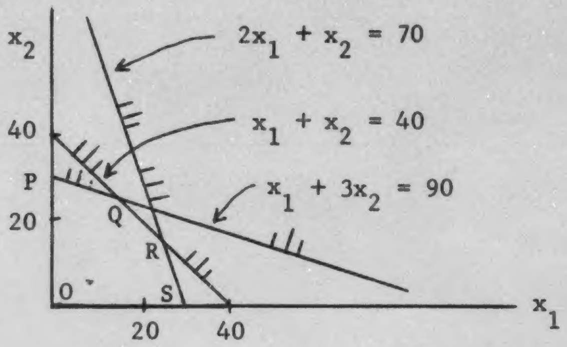
(6) Repeat steps (3)-(5) until there are no negative elements in the last row.

7.3 Geometrical approach

The inequality $x_1 + x_2 \leqq 40$ of the preceding example can be represented graphically as follows:

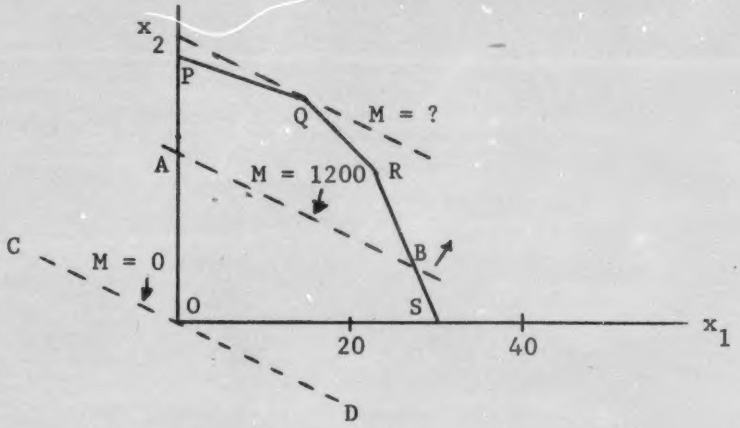


The shading indicates that points (x_1,x_2) satisfying the inequality lie below the line. The complete diagram for all the inequalities of (2) is:



The set of feasible points lies inside or on the polygon $\overline{OPQRS}$.

In the next diagram we add the lines $M = 40x_1 + 60x_2$ for $M = 0$ (the line $\overline{CD}$) and $M = 1200$ (the line $\overline{AB}$).



All points on the line segment $\overline{AB}$ in the diagram are feasible, and they correspond to M = 1200. If we move the line $\overline{AB}$ parallel to itself (in the direction of the arrow), we still have a line of the form $M = 40x_1 + 60x_2$. The line which maximizes M is the one through Q. The point Q is the intersection of

$$x_1 + x_2 = 40 \text{ and } x_1 + 3x_2 = 90,$$

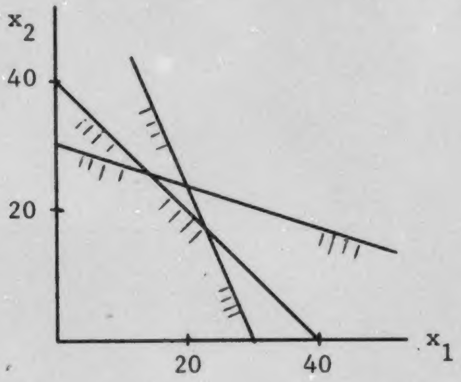
i.e.,

$$x_1 = 15,\ x_2 = 25, \text{ and hence } M = 2100.$$

These are the values obtained in the last section algebraically.

There are two advantages of the geometrical approach:

(1) We can easily see the kinds of complications that may arise. For example, if, instead of passing through a single vertex Q, the optimum line coincides with a side of the polygon, there will be an infinity of solutions. If the original problem involved the inequalities $2x_1 + x_2 \geqq 70$, $x_1 + 3x_2 \geqq 90$, $x_1 + x_2 \leqq 40$ instead of those given previously, we would have:



Here it is clear that there is no solution.

(2) It is possible to see how to generalize to n dimensions by introducing the "line" and "hyperplane" in n dimensions. A convex set is a set with the property that whenever $\underline{x}_1$ and $\underline{x}_2$ are in the set, then so is $\lambda\underline{x}_1 + (1 - \lambda)\underline{x}_2$, if $0 \leqq \lambda \leqq 1$. It may be proved that the set of feasible solutions is a convex set. The simplex method corresponds to proceeding from one extreme point to an adjacent extreme point along an edge of the set until the maximum is determined.

7.4 Theory of games

Consider the following game between two players A and B. Each has a coin which he may turn up heads or tails without the knowledge of the other player. The coins are then compared and A receives a "payoff" according to the faces of the coins turned up, as shown in the table.

A	B	Payoff to A
H	H	1¢
H	T	-2¢
T	H	-3¢
T	T	4¢

A negative payoff to A is to be interpreted as a gain for B. Each player must decide what percentage of the time to choose heads and what percentage to choose tails in order to maximize his gain.

It is assumed that each player will choose randomly. (If a pattern is followed, great losses could be inflicted if the opponent guesses the pattern.) Thus, if A chooses heads with probability x and B chooses heads with probability y, then the average payoff to A will be

$$\begin{aligned} P &= x[y - 2(1 - y)] + (1 - x)[-3y + 4(1 - y)] \\ &= 10\,(x - .7)(y - .6) - 0.2\,. \end{aligned} \tag{14}$$

If A chooses x to be .7, then P = -0.2 *independently* of what B does. If he chooses any other value of x, a good guess of y by B might lead to even greater losses. Thus x = .7 is the best strategy available to A. On the other hand, B can *insure* himself a gain of 0.2 by choosing y = .6. Although another choice might lead to greater return, it could also lead to loss if A plays properly. Thus y = .6 is the optimum strategy for B. The preceding table may be conveniently summarized in a *payoff matrix*

$$\begin{array}{ccc} & & \text{B} \\ & & \begin{array}{cc} \text{H} & \quad\text{T} \end{array} \\ \text{A} & \begin{array}{c}\text{H}\\ \text{T}\end{array} & \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \end{array}$$

which facilitates the analysis in more complicated cases.

We will show how linear programming can be used to solve a somewhat more difficult example. Suppose the payoff matrix is

$$\begin{array}{cc} & & \text{B} & \\ & & s & t \\ & p & \left[\begin{array}{c} 2 \\ 3 \\ 0 \end{array}\right. & \left.\begin{array}{c} 5 \\ 1 \\ 3 \end{array}\right] \\ \text{A} & q & & \\ & r & & \end{array}$$

$$\begin{array}{ccc} & s & t \\ p & 2 & 5 \\ q & 3 & 1 \\ r & 0 & 3 \end{array}$$

where p, q, r are the probabilities to be used by A and s, t are the probabilities to be used by B. Thus $p + q + r = 1$, $s + t = 1$. Let g be the smaller of the quantities $2p + 3q$ and $5p + q + 3r$. The payoff to A will be

$$s(2p + 3q) + t(5p + q + 3r) \geqq sg + tg = g \qquad (15)$$

The payoff to A is therefore at least g, no matter what B does. Thus A wishes to choose p, q, and r to maximize g.

Since A can clearly do better than $g = 0$, we assume $g > 0$ and introduce $x_1 = \frac{p}{g}$, $x_2 = \frac{q}{g}$, and $x_3 = \frac{r}{g}$. The conditions of the game become:

$$\begin{aligned} 2x_1 + 3x_2 \qquad &\geqq 1 \\ 5x_1 + x_2 + 3x_3 &\geqq 1 \\ x_1 + x_2 + x_3 &\geqq \frac{1}{g} \\ x_1 \geqq 0,\ x_2 \geqq 0,\ x_3 &\geqq 0 \end{aligned} \qquad (16)$$

Thus A seeks to minimize $x_1 + x_2 + x_3$, a problem which is clearly amenable to linear programming methods.

The problem, from the point of view of B, is to minimize the payoff to A. Thus B wishes to minimize the largest of $2s + 5t$, $3s + t$, and $3t$. If h is the largest of these quantities, B formulates the problem by introducing $y_1 = \frac{s}{h}$, $y_2 = \frac{t}{h}$ and seeking to maximize $y_1 + y_2$ subject to:

$$2y_1 + 5y_2 \leqq 1$$

$$3y_1 + y_2 \leqq 1$$

$$3y_2 \leqq 1$$

$$y_1 \geqq 0,\ y_2 \geqq 0 \tag{17}$$

The linear programming problems that arise from these two points of view are clearly related in some way. The elucidation of this relationship introduces the idea of duality. If the primal problem is to maximize

$$\underline{c}^T\underline{x},\ \text{with}\ \underline{x} \geqq 0,\ A\underline{x} \leqq \underline{b}, \tag{18}$$

then the dual problem is to minimize

$$\underline{b}^T\underline{y},\ \text{with}\ \underline{y} \geqq 0,\ A^T\underline{y} \geqq \underline{c}. \tag{19}$$

The connection proves to be that if (18) has a solution $\underline{x}$, then (19) has a solution $\underline{y}$ and

$$\underline{c}^T\underline{x} = \underline{b}^T\underline{y}. \tag{20}$$

It is easy to show that $\underline{c}^T\underline{x} \leq \underline{b}^T\underline{y}$ as follows:

$$A^T\underline{y} \geqq \underline{c}\ \text{implies}\ \underline{y}^TA \geqq \underline{c}^T,$$

so

$$\underline{c}^T\underline{x} \leq \underline{y}^TA\underline{x} \leq \underline{y}^T\underline{b}. \tag{21}$$

The complete proof is more difficult.

The main points of this section are that (1) the theory of games may be formulated in linear programming terms, and (2) the respective viewpoints of two players illustrate the primal and dual problems of linear programming.

Many books on linear programming and the theory of games have been published. An interesting, simple, but nontrivial treatment is given by Glicksman [22]. See also Noble [43].

LINEAR ALGEBRA

Lecture 8: Application of Matrices to Economic Models and Social Science Relationships

Mathematical topics: Matrix multiplication, existence of inverses and solutions of linear systems when the matrices involved have special properties

Comment: Formulation of certain problems in economics leads naturally to linear equations, both of homogeneous and inhomogeneous type. In either case it is found that the restrictions placed on the matrix elements by economic considerations lead to physically reasonable (nonnegative) solutions. It is possible to express certain social relationships in terms of matrices the powers of which give information about cliques, intercommunication between individuals, etc.

8.1 Input-output (Leontief closed) models

Consider a simple society consisting of three individuals, a farmer F who produces all the food, a builder B who constructs all the shelter, and a tailor T who makes all the clothing. Suppose the fraction of each commodity consumed by each of the individuals is given by the following table.

	Food	Shelter	Clothing
Farmer	7/16	1/2	3/16
Builder	5/16	1/6	5/16
Tailor	1/4	1/3	1/2

Suppose that the income to the farmer (F) is p_1, to the builder (B) is p_2 and to the tailor (T) is p_3. It is desired to determine whether the facts given in the table are compatible with the equilibrium condition that consumption is equal to production.

Consider that each sells to and buys from a central pool. Assuming equilibrium, it is evident that

$$\frac{7}{16} p_1 + \frac{1}{2} p_2 + \frac{3}{16} p_3 = p_1$$

$$\frac{5}{16} p_1 + \frac{1}{6} p_2 + \frac{5}{16} p_3 = p_2$$

$$\frac{1}{4} p_1 + \frac{1}{3} p_2 + \frac{1}{2} p_3 = p_3. \tag{1}$$

This leads to the homogeneous system

$$(I - A)\, \underline{p} = 0, \tag{2}$$

where

$$\underline{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}, \quad I - A = \begin{bmatrix} \frac{9}{16} & -\frac{1}{2} & -\frac{3}{16} \\ -\frac{5}{16} & \frac{5}{6} & -\frac{5}{16} \\ -\frac{1}{4} & -\frac{1}{3} & \frac{1}{2} \end{bmatrix}.$$

The column sums of I - A are all zero, so I - A is singular and there is an infinity of solutions. The solutions are

$$\underline{p} = \alpha \begin{bmatrix} 4 \\ 3 \\ 4 \end{bmatrix}. \tag{4}$$

Thus the incomes of F, B, and T are in the ratios 4:3:4, respectively.

The model is readily extended. If the system includes a laborer L, the matrix A might be:

	Food	Shelter	Clothing	Labor
F	1/3	1/3	1/4	1/3
B	1/4	1/3	5/16	1/6
T	1/4	1/6	3/16	1/4
L	1/6	1/6	1/4	1/4

The matrix A is called the exchange or input-output matrix. In the general case of n parties (which may be thought of as individuals, companies, industries, etc.) the matrix element a_{ij} is the fraction of goods produced by j which are consumed by i.

The Leontief model makes the reasonable assumptions

$$0 \leqq a_{ij} \leqq 1$$

$$\sum_i a_{ij} = 1, \quad j = 1, \ldots, n,$$

with the equilibrium condition

$$A\underline{p} = \underline{p}. \tag{5}$$

It is desired to know whether solutions of (5) other than $\underline{p} = 0$ will always exist and whether they are physically reasonable, i.e., whether $\underline{p} \geqq 0$.

We need a preliminary result.

Theorem 8.1. If S is an n × m matrix of elements $s_{ij} \geqq 0$ and the column sums of S are less than unity, then I - S is nonsingular and the elements of $(I - S)^{-1}$ are nonnegative.

Proof: Define the norms

$$\|x\| = \sum_{i=1}^{n} |x_i|$$

$$\|S\| = \max_j \sum_{i=1}^{n} |s_{ij}|. \tag{6}$$

By hypothesis $\|S\| < 1$, so the existence of $(I - S)^{-1}$ follows from Theorem 4.3. Also, since

$$(I - S)^{-1} = I + S + S^2 + \ldots \tag{7}$$

it is clear that no element of $(I - S)^{-1}$ is negative.

As the next step we prove

Theorem 8.2. If $A = \begin{bmatrix} A_{11} & \underline{a} \\ \underline{b}^T & \alpha \end{bmatrix}$, where $\underline{b} > 0$, $\underline{a} > 0$, $0 \leqq a_{ij} \leqq 1$ for each i,j, and $\sum_i a_{ij} = 1$ for each j, then $(I - A)\,\underline{x} = 0$ has a one-dimensional solution space of vectors

$$\underline{x} = \beta\underline{p}, \text{ where } \underline{p} \geqq 0 \text{ and } p_n > 0.$$

121

Proof: First notice that the column sums of I - A are all zero, so I - A is singular and the system $(I - A)\underline{x} = 0$ has infinitely many solutions. For some nontrivial solution $\underline{x}$ we write $\underline{x} = \begin{bmatrix} \underline{y} \\ x_n \end{bmatrix}$ and $I = \begin{bmatrix} I_1 & 0 \\ 0 & 1 \end{bmatrix}$ where $\underline{y}$ is an $(n - 1) \times 1$ matrix and I_1 is the $(n - 1) \times (n - 1)$ unit matrix. Then, by hypothesis,

$$0 = (I - A)\underline{x} = \begin{bmatrix} I_1 - A_{11} & -\underline{a} \\ -\underline{b}^T & 1 - \alpha \end{bmatrix} \begin{bmatrix} \underline{y} \\ x_n \end{bmatrix}, \qquad (8)$$

which implies that

$$(I_1 - A_{11})\underline{y} - x_n\underline{a} = 0 \qquad (9)$$

and

$$-\underline{b}^T\underline{y} + (1 - \alpha)x_n = 0.$$

Since the column sums of A are all unity and since $\underline{b} > 0$, the column sums of A_{11} are all less than unity. Therefore, by Theorem 8.1, $(I_1 - A_{11})^{-1}$ exists and has no negative elements. Thus we may solve (9) for $\underline{y}$:

$$\underline{y} = x_n (I_1 - A_{11})^{-1}\underline{a} \qquad (10)$$

If $x_n = 0$, then $\underline{y} = 0$ and $\underline{x}$ would be the trivial solution. Consequently, we may write

$$\underline{x} = x_n \begin{bmatrix} (I_1 - A_{11})^{-1}\underline{a} \\ 1 \end{bmatrix}, \qquad (11)$$

$$= x_n\underline{p}.$$

It is clear that $\underline{p}$ is positive and independent of $\underline{x}$, so the proof is complete.

The system (5) is precisely a homogeneous system $(I - A)\underline{p} = 0$, fulfilling the hypotheses of Theorem 8.2. Thus there are always nontrivial, physically reasonable solutions. As was seen in the example, the constant β does not affect the ratios of the incomes and may be thought of as establishing the unit of value for the currency.

At first sight it might seem reasonable to employ $A\underline{p} \leqq \underline{p}$ as a more flexible equilibrium condition. However, it is easily seen that $A\underline{p} \leqq \underline{p}$ in fact implies $A\underline{p} = \underline{p}$. Suppose $A\underline{p} \leqq \underline{p}$. Then

$$\sum_j a_{ij}p_j < p_i, \quad i = 1, \ldots, n, \tag{12}$$

and, summing over i,

$$\begin{aligned} \sum_i p_i &> \sum_i \sum_j a_{ij}p_j \\ &> \sum_j p_j \sum_i a_{ij} \\ &> \sum_j p_j \end{aligned} \tag{13}$$

since the column sums of A are all one. But this is contradictory, so $A\underline{p}$ is not less than $\underline{p}$.

In Theorem 8.2 we have seen conditions under which there is only a one-parameter family of solutions. The hypothesis is easily verified if $A > 0$. If A contains zeros, however, the hypothesis may not be fulfilled. As an example, suppose

$$A = \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{2} \\ \frac{2}{3} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}.$$

The solution of $A\underline{p} = \underline{p}$ is

$$\underline{p} = \alpha \begin{bmatrix} 3 \\ 0 \\ 4 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}. \tag{15}$$

The economic interpretation is that there are two disjoint sets of industries.

The situation may be clarified by introducing the idea of reducibility. A matrix A is reducible if it can be brought to the form

$$A = \begin{bmatrix} P & Q \\ 0 & S \end{bmatrix} \tag{16}$$

by any reordering of the rows together with the same reordering of the columns. If A is reducible, our problem is

$$\begin{bmatrix} P & Q \\ 0 & S \end{bmatrix} \begin{bmatrix} \underline{y} \\ \underline{z} \end{bmatrix} = \begin{bmatrix} P\underline{y} + Q\underline{z} \\ S\underline{z} \end{bmatrix} = \begin{bmatrix} \underline{y} \\ \underline{z} \end{bmatrix} \tag{17}$$

or

$$\begin{aligned} P\underline{y} + Q\underline{z} &= \underline{y} \\ S\underline{z} &= \underline{z} \, . \end{aligned} \tag{18}$$

If nontrivial $\underline{y}$ and $\underline{z}$ satisfy (18), then the linearly independent vectors $\begin{bmatrix} \underline{y} \\ \underline{z} \end{bmatrix}$ and $\begin{bmatrix} \underline{y} \\ \underline{0} \end{bmatrix}$ are solutions to the original problem.

8.3 The production (Leontief open) model

In the open model it is assumed that for good j there is an outside demand d_j in addition to the amount a_{ij} of good j consumed in the production of good i. If x_i is the amount of good i produced, then the equilibrium condition is that

$$x_j - \sum_i x_i a_{ij} = d_j, \text{ for each } j, \tag{19}$$

or, in matrix notation,

$$\underline{x}^T(I - A) = \underline{d}^T, \tag{20}$$

subject to

$$a_{ij} \geqq 0, \; d_j \geqq 0. \tag{21}$$

The problem, as before, is to investigate the existence of a solution $\underline{x}$ of (20) which is physically reasonable, i.e., $\underline{x} \geqq 0$.

The central issue to be settled is clearly the existence of $(I - A)^{-1}$. If this matrix exists, then

$$\begin{aligned} \underline{x}^T &= \underline{d}^T (I - A)^{-1} \\ &= \underline{d}^T (I + A + A^2 + \ldots). \end{aligned} \tag{22}$$

The latter expression is manifestly $\geqq 0$ since $\underline{d}$ and A have no negative elements. Several theorems might be applied for special forms of A. We consider only one such case.

If $A \geqq 0$ and there exists any vector $\underline{b} > 0$ such that $\underline{x}^T (I - A) = \underline{b}^T$ has a solution $\underline{x} \geqq 0$, then A is said to be productive. The following theorem shows that the problem can be solved when A is productive:

Theorem 8.3. If A is productive, then $I + A + A^2 + \ldots$ converges.

Proof: By hypothesis, there exist $\underline{x} \geqq 0$ and $\underline{b} > 0$ such that

$$\underline{x}^T (I - A) = \underline{b},$$

so that

$$\underline{x}^T (I - A) > 0$$

or

$$\underline{x}^T > \underline{x}^T A \geqq 0.$$

Thus for some λ $(0 < \lambda < 1)$, $\lambda \underline{x}^T > \underline{x}^T A$. Consequently $\lambda^n \underline{x}^T > \underline{x}^T A^n$, which approaches zero as n increases, since $0 < \lambda < 1$. Now this is possible for positive $\underline{x}$ and A only if $A^n \to 0$ as $n \to \infty$. This is sufficient for the convergence of the matrix geometric series $I + A + A^2 + \ldots$. Thus we are able to assert that $(I - A)^{-1}$ exists and that all of its elements are positive.

The point of the Leontief model for teaching is to show the interaction of economics and mathematics. We write down mathematical equations for economic behavior. The physical interpretation places restrictions on the matrix elements in these equations. These restrictions are seen to be sufficient to guarantee that the solutions of the mathematical problem have a physically reasonable meaning.

8.4 Social sciences

We will indicate briefly some uses of matrices in the social sciences. Sources listed in the bibliography may be consulted for additional details.

Consider a group of five schoolgirls, each of whom wishes to be friends with some of the others. We can represent the relationship "wants to be friends" in a matrix notation by the following scheme: a one in the (i,j) position means that girl i wishes to be friends with girl j; a zero means that she does not. We suppose the given relationships are summarized in the following matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}. \qquad (23)$$

This is an example of a general class of relationships in which the relationship of i to j is independent of the relationship of j to i. In the above case, for example, girl 1 has the relationship (i.e., wants to be friends) with girl 2, but girl 2 does not reciprocate. Notice that we have assumed that the relation is not to be regarded as possible between a girl and herself, as reflected by the zeros on the diagonal.

The matrix multiplication law has a ready interpretation which we state as a theorem.

Theorem 8.4. Let A be the matrix of a relation and let A^2 have elements $a_{ij}^{(2)}$. Then $a_{ij}^{(2)}$ is the number of distinct, two-step connections between i and j.

Proof: Since $a_{ij}^{(2)} = \sum_{k=1}^{n} a_{ik} a_{kj}$, the result is obvious.

A clique is defined to be a collection of three or more persons each of whom possesses the given relationship with each other member of the collection, and such that the collection is contained in no larger set with this property. The existence of cliques may also be investigated by use of matrix operations. Define a matrix S as follows:

$$s_{ij} = \begin{cases} 1 \text{ if } a_{ij} \text{ and } a_{ji} \text{ are } 1 \\ 0 \text{ otherwise} \end{cases} \qquad (24)$$

The following theorems may now be used to identify cliques.

Theorem 8.5. Let S^3 have elements $s_{ij}^{(3)}$. Then i is a member of a clique if and only if $s_{ii}^{(3)} > 0$.

Theorem 8.6. If the i^{th} person is a member of exactly one clique, then this is an m-member clique if $s_{ii}^{(3)} = (m - 1)(m - 2)$.

A relation is called a dominance relation if, for each two persons, exactly one of them possesses the relation to the other. Using the same conventions as before, the following is the matrix of a dominance relation:

$$B = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \tag{25}$$

The following theorem can be established by matrix methods.

Theorem 8.7. If a dominance relation holds in a group, then (a) there is one member of the group in either a one-step or two-step relationship to every other member of the group, and (b) there is one member to whom every other member is in a one- or two-step relationship.

Mathematically this states that $B + B^2$ has one row which contains ones in every positions except on the diagonal, and one column which contains ones in every position except on the diagonal.

An intrabusiness relation is one in which all relationships are reciprocated. The matrix of the relationships is therefore symmetric. In this case we put ones on the diagonal and can prove the following theorem.

Theorem 8.8. Let A be the matrix for an intrabusiness relation. If $A + A^2 + \ldots + A^{n-1}$ has a zero entry, then there is a subset of the group no member of which has a relationship with anyone outside the subset.

8.5 Notes and references

Elementary treatments of Leontief models are given in Boot [8], Johnston, et al. [34], and Yan [52]. Allen [1] may also be consulted, and, for a more advanced treatment, Gale [19]. The treatment of social relationships in 8.4 was taken directly from Johnston, et al. [34], Chapter 4, where proofs and references to original work can be found.

Lecture 9: Linear Least Squares Method; Regression

Mathematical topics: Linear equations, formal differentiation with respect to a vector, idempotent matrices

Comment: The importance of techniques for fitting data points with a mathematical model will be evident to students in many disciplines. The linear model is treated in the first section; this is followed by an analysis of the statistical significance of the results obtained. Both topics rely heavily on matrix methods.

9.1 Linear Regression Analysis; Least Squares Method

The simplest example of linear regression is the problem of fitting a straight line to a set of data points in two-dimensional space. Suppose that n points (x_1,y_1), (x_2,y_2), ..., (x_n,y_n) are given, where $x_1 < x_2 < \ldots < x_n$. Find the line $y = b_o + b_1x$ which "best" fits the n points. This is done by specifying the coefficients b_o and b_1. If the student is given several specific points and asked to find the "best" fitting line for these points, he will soon realize that some criterion must be set up for determining the "best" fitting line.

With b_o and b_1 still unknown, we write $y_i = b_o + b_1x_i + r_i$, $i = 1, \ldots, n$. The term r_i is called the i^{th} residual and is the error in approximating y_i by $b_o + b_1x_i$. Geometrically it is the vertical directed distance from the point (x_i,y_i) to the line $y = b_o + b_1x$. Ideally there should be no error in the x_i values, but this is usually not the case. There are two causes for error in the approximation of y_i by $b_o + b_1x_i$. First, there will always be error in the data due to error in measurements. Second, the linear equation may not be a good model for the given set of data. For instance, an exponential curve might fit the data points much better than a line.

Let S equal the sum of the squares of the residuals, that is, $S = \sum_{i=1}^{n} r_i^2$. Associated with each line in the plane is an S value; the "best" fitting line is chosen to be the one with smallest S value (note $S \geqq 0$). This is the criterion for choosing the "best" fitting line; it is called the least squares method.

At this point the student may ask why $\sum_{i=1}^{n} r_i$ would not be a suitable choice for S. It is easy to show that this choice would give negative S values and that more than one line could have zero S value. For a simple example consider the two points (0,0) and (2,2). The line $y = x$ has $r_1 + r_2 = 0 + 0 = 0$ and the line $y = 1$ has $r_1 + r_2 = -1 + 1 = 0$. It is obvious that $y = x$ is the best fitting line, but choosing S equal to $\sum_{i=1}^{n} r_i$ would not indicate this.

There are two other possible choices for S. One is $S = \sum_{i=1}^{n} |r_i|$ and in the other S is equal to the sum of the perpendicular distances from the data points to the line $y = b_o + b_1 x$. These are both reasonable choices for S; however, both involve absolute values. This would make $\frac{\partial S}{\partial b_o}$ and $\frac{\partial S}{\partial b_1}$ impossible to simplify. Thus S cannot be minimized in these cases even though unique lines with minimum S values probably do exist.

It is interesting to note here that finding the "best" fitting line, that is, finding b_o and b_1, can be done using a property of certain quadratic equations. An outline of this derivation is as follows.

$$S = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - b_o - b_1 x_i)^2$$

If this is expanded, we obtain an expression for S which is quadratic in b_o, treating b_1 as a constant, with positive leading coefficient. Also, S can be expressed as a quadratic in b_1, treating b_o as a constant, with positive leading coefficient. Now we use the fact that the dependent variable of a quadratic equation with a positive leading coefficient is minimized at the vertex. From the two quadratic equations described above we can obtain a system of two linear equations with unknowns b_o and b_1. The solution of this system is given by

$$b_o = \frac{\left(\Sigma y_i\right)\left(\Sigma x_i^2\right) - \left(\Sigma x_i\right)\left(\Sigma y_i x_i\right)}{n\left(\Sigma x_i^2\right) - \left(\Sigma x_i\right)^2}$$

$$b_1 = \frac{n\left(\Sigma y_i x_i\right) - \left(\Sigma y_i\right)\left(\Sigma x_i\right)}{n\left(\Sigma x_i^2\right) - \left(\Sigma x_i\right)^2} .$$

(Here every summation is from 1 to n.) Thus b_o and b_1 are uniquely determined.

The above derivation could be incorporated into a college algebra course. It would be very useful to the student in laboratory courses, and it would provide the linear algebra student with an early introduction to the least squares method. Unfortunately, this derivation does not give insight into the solution of the problem when more than two unknowns are needed.

The matrix method of solving for the "best" fitting line is as follows. The system of equations

$$\begin{aligned} y_1 &= b_o + b_1 x_1 + r_1 \\ y_2 &= b_o + b_1 x_2 + r_2 \\ &\vdots \qquad \vdots \\ y_n &= b_o + b_1 x_n + r_n \end{aligned}$$

can be written as $\underline{y} = X\underline{b} + \underline{r}$, where

$$\underline{y} = [y_1, y_2, \ldots, y_n]^T, \quad \underline{r} = [r_1, r_2, \ldots, r_n]^T,$$

$$\underline{b} = \begin{bmatrix} b_o \\ b_1 \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix}^T .$$

Now

$$\begin{aligned} S = \sum_{i=1}^{n} r_i^2 &= \underline{r}^T\underline{r} = (\underline{y} - X\underline{b})^T(\underline{y} - X\underline{b}) \\ &= (\underline{y}^T - \underline{b}^T X^T)(\underline{y} - X\underline{b}) \\ &= \underline{y}^T\underline{y} - \underline{y}^T X\underline{b} - \underline{b}^T X^T \underline{y} + \underline{b}^T X^T X\underline{b} \\ &= \underline{y}^T\underline{y} - 2\underline{y}^T X\underline{b} + \underline{b}^T X^T X\underline{b}. \end{aligned}$$

It is an easy exercise to show that the scalars $\underline{y}^T X \underline{b}$ and $\underline{b}^T X^T \underline{y}$ are equal. S is a scalar function of b_o and b_1. S is minimized by finding b_o and b_1 which make $\frac{\partial S}{\partial b_o} = 0$ and $\frac{\partial S}{\partial b_1} = 0$. However, it is more instructive at this point to introduce a partial derivative with respect to a vector of variables. This approach makes the more general least squares problem easy to solve.

Let $\lambda = a_1 x_1 + \ldots + a_n x_n = [a_1 \ldots a_n] \begin{bmatrix} x_1 \\ \cdot \\ \vdots \\ x_n \end{bmatrix} = \underline{a}^T \underline{x}$, where

$\underline{a} = \begin{bmatrix} a_1 \\ \cdot \\ \vdots \\ a_n \end{bmatrix}$ and $\underline{x}$ consists of variables.

Define $$\frac{\partial \lambda}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial \lambda}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial \lambda}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial}{\partial x_n} \end{bmatrix} \lambda .$$

Theorem. If $\lambda = \underline{a}^T \underline{x}$, then $\frac{\partial \lambda}{\partial \underline{x}} = \underline{a}$.

Proof: Immediate from the definition.

Let $\underline{y} = A\underline{x}$, where $\underline{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix}$, $\underline{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$, and $A = (a_{ij})_{m \times n}$.

Define $$\frac{\partial \underline{y}}{\partial \underline{x}} = \left[\frac{\partial y_1}{\partial \underline{x}}, \frac{\partial y_2}{\partial \underline{x}}, \ldots, \frac{\partial y_m}{\partial \underline{x}} \right].$$

Note that $$\frac{\partial \underline{y}}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}.$$

Theorem. If $\underline{y} = A\underline{x}$, then $\frac{\partial \underline{y}}{\partial \underline{x}} = A^T$.

Proof: If a_i denotes the i^{th} row of A, then

$$\begin{aligned}
\frac{\partial \underline{y}}{\partial \underline{x}} &= \left[\frac{\partial y_1}{\partial \underline{x}}, \ldots, \frac{\partial y_m}{\partial \underline{x}} \right] \\
&= \left[\frac{\partial (a_1 \underline{x})}{\partial \underline{x}}, \ldots, \frac{\partial (a_m \underline{x})}{\partial \underline{x}} \right] \\
&= \left[a_1^T, \ldots, a_m^T \right] \\
&= A^T.
\end{aligned}$$

Now let $A = (a_{ij})_{n \times n}$.

Theorem. If $y = \underline{x}^T A\underline{x}$, then $\frac{\partial y}{\partial \underline{x}} = A\underline{x} + A^T\underline{x}$.

Proof: If the a_i's are the rows of A and the α_i's are the columns of A, then

$$\frac{\partial y}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial}{\partial x_n} \end{bmatrix} \left(\sum_j \sum_i a_{ij} x_i x_j \right) = \begin{bmatrix} \sum_{j=1}^{n} a_{1j} x_j + \sum_{i=1}^{n} a_{i1} x_i \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{n} a_{nj} x_j + \sum_{i=1}^{n} a_{in} x_i \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \underline{x} + \alpha_1^T \underline{x} \\ \cdot \\ \cdot \\ \cdot \\ a_n \underline{x} + \alpha_n^T \underline{x} \end{bmatrix}$$

$$= A\underline{x} + A^T\underline{x}.$$

Corollary. If $A = A^T$, then $\frac{\partial (\underline{x}^T A \underline{x})}{\partial \underline{x}} = 2A\underline{x}$.

Recall that $S = \underline{y}^T\underline{y} - 2\underline{y}^TX\underline{b} + \underline{b}^TX^TX\underline{b}$. Now,

$$\frac{\partial S}{\partial \underline{b}} = -2(\underline{y}^TX)^T + 2(X^TX)\underline{b}$$

$$= -2X^T\underline{y} + 2X^TX\underline{b},$$

using the above theorem for $\frac{\partial}{\partial \underline{b}}$. (Notice that X^TX is symmetric.) S is minimized when $\frac{\partial S}{\partial \underline{b}} = 0$, that is, when $X^TX\underline{b} = X^T\underline{y}$ or $\underline{b} = (X^TX)^{-1}\ X^T\underline{y}$.

$$X^TX = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \qquad X^T\underline{y} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

$$\left(X^TX\right)^{-1} = \frac{1}{n\Sigma x_i^2 - \left(\Sigma x_i\right)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$$

Thus,

$$\underline{b} = \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \frac{1}{n\Sigma x_i^2 - \left(\Sigma x_i\right)^2} \begin{bmatrix} \Sigma y_i\ \Sigma x_i^2 - \Sigma x_i\ \Sigma x_i y_i \\ -\Sigma x_i\ \Sigma y_i + n\Sigma x_i y_i \end{bmatrix}.$$

As before, all summations are from 1 to n.

At this point an example would be helpful to the student. Schaum's outline in Numerical Analysis gives some good examples. An in-class experiment in which the least squares method is used to interpret the data would be particularly illustrative. Estimating a spring constant would be one such experiment that could easily be performed.

The following economic example is taken from Searle and Hausman [49], page 251. Let y_i = rate of inflation during time period i and let x_i = rate of unemployment during period i. A model is sought whereby the rate of inflation can be predicted by the rate of unemployment. One possible model is $y = b_o + b_1x$.

Data of the example

i	y_i	x_i
1	2.0	5.0
2	3.0	4.2
3	3.2	4.3
4	4.0	3.8
5	3.6	4.4
6	2.1	5.2
7	1.7	6.0
8	1.2	6.5
9	1.8	5.9
10	2.4	4.0

$$\sum_{1}^{10} y_i = 25.0 \qquad \sum_{1}^{10} x_i = 49.3$$

$$\sum_{1}^{10} y_i^2 = 69.94 \qquad \sum_{1}^{10} x_i^2 = 251.03$$

$$\sum_{1}^{10} y_i x_i = 116.54$$

Substituting into the matrix equation for $\underline{b}$ gives

$$\underline{b} = \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \begin{bmatrix} 6.65 \\ -0.841 \end{bmatrix} .$$

Thus the "best" fitting line for the data of this problem is $y = 6.65 - 0.841\ x$.

The more general situation can now be solved with ease. Suppose that associated with each k-tuple $(x_{i1}, x_{i2}, \ldots, x_{ik})$ is a real number y_i, where $i = 1, 2, \ldots, n$ and $n > k$. Find an equation of the form $y = b_o + b_1x_1 + \ldots + b_kx_k$ which "best" describes the given data. A system of n equations with k + 1 unknowns $b_o, \ldots, b_k$ can be written as

$$\{y_i = b_o + b_1 x_{i1} + \ldots + b_k x_{ik} + r_i;\ i = 1, \ldots, n\}.$$

Each r_i is the error in approximating y_i by $b_o + b_1 x_{i1} + \ldots + b_k x_{ik}$ and is again called the i^{th} residual. As before, let $S = \sum_{i=1}^{n} r_i^2$ and agree that the equation $y = b_o + b_1 x_1 + \ldots + b_k x_k$ which "best" describes the data is the one with the smallest value for S.

The system of n equations given above can be written as $\underline{y} = X\underline{b} + \underline{r}$, where $\underline{y} = [y_1, \ldots, y_n]^T$, $\underline{b} = [b_o, \ldots, b_k]$, $\underline{r} = [r_1, \ldots, r_n]^T$, and

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad (n > k).$$

The matrix equation $\underline{y} = X\underline{b} + \underline{r}$ is exactly the same as in the two-dimensional situation. Moreover, the matrix operations required to solve for $\underline{b}$ are exactly the same. Thus, as before,

$$S = \sum_{i=1}^{n} r_i^2 = \underline{r}^T\underline{r} = (\underline{y} - X\underline{b})^T(\underline{y} - X\underline{b})$$

$$= \underline{y}^T\underline{y} - 2\underline{y}^T X\underline{b} + \underline{b}^T X^T X\underline{b}$$

and $\frac{\partial S}{\partial b} = -2X^T\underline{y} + 2X^T X\underline{b}$. S is minimized when $\frac{\partial S}{\partial \underline{b}} = 0$. Thus $X^T X\underline{b} = X^T\underline{y}$ or $\underline{b} = (X^T X)^{-1} X^T \underline{y}$ minimizes S, provided that $(X^T X)^{-1}$ exists. The matrix X is an n × (k+1) matrix and $n \geqq k + 1$. If the rank of X is equal to k + 1, then $(X^T X)^{-1}$ exists. This statement is proved on page 139 of Noble [43].

9.2 Accuracy of the estimates of $b_o, b_1, \ldots, b_n$

Our point of view is that there exists a function

$$y(\underline{x}) = b_o + \sum_i b_i x_i \qquad (1)$$

which would exactly represent the given data if these data were free of error. In the last section an estimate

$$\underline{\hat{b}} = (X^T X)^{-1} X^T \underline{y} \tag{2}$$

was obtained for the coefficient vector

$$\underline{b} = [b_o, b_1, \ldots, b_n]^T \tag{3}$$

from some given set of data which presumably contained observational uncertainty. It is therefore a problem of statistics to determine the accuracy of the estimates $\underline{\hat{b}}$ of the quantity $\underline{b}$. Solution of this problem provides another useful application of matrix methods.

We recall some concepts from elementary statistics. Let $\underline{z} = [z_1, \ldots, z_n]^T$ be the vector of n observations of the quantity z. The expected value of z, denoted E(z), is

$$E(z) = \frac{1}{n} \sum_{i=1}^{n} z_i . \tag{4}$$

The variance of z, denoted var(z), is

$$\mathrm{var}(z) = E[(z - \overline{z})^2] = \frac{1}{n} \sum_{i=1}^{n} (z_i - \overline{z})^2 . \tag{5}$$

The covariance of the two variables y and z, denoted cov(y,z) is

$$\begin{aligned} \mathrm{cov}(y,z) &= E[(y - \overline{y})(z - \overline{z})] \\ &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(z_i - \overline{z}) \\ &= \mathrm{cov}(z,y). \end{aligned} \tag{6}$$

Note that cov(z,z) = var(z). For uncorrelated variables y and z, we would expect cov(y,z) to be small, since the terms should occur with negative signs about as often as with positive signs. For a set of variables $v_1, v_2, \ldots, v_n$, the variance-covariance matrix is

$$\begin{bmatrix} \mathrm{var}(v_1) & \mathrm{cov}(v_1,v_2) & \cdots & \mathrm{cov}(v_1,v_n) \\ \mathrm{cov}(v_2,v_1) & \mathrm{var}(v_2) & \cdots & \mathrm{cov}(v_2,v_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(v_n,v_1) & \mathrm{cov}(v_n,v_2) & \cdots & \mathrm{var}(v_n) \end{bmatrix}. \qquad (7)$$

This is denoted $E[(\underline{v} - \overline{\underline{v}})(\underline{v} - \overline{\underline{v}}^T)]$, where $\underline{v} = [v_1, \ldots, v_n]^T$.

The exact vector $X\underline{b}$ is not identical with the observed vector $\underline{y}$, so we define the remainder $\underline{r}$ by the equation

$$\underline{y} = X\underline{b} + \underline{r}. \qquad (8)$$

Since the observational error responsible for $\underline{r}$ is presumably random, we assume that if $\underline{r}$ is determined from a large number of data sets, then

$$E(\underline{r}) = 0. \qquad (9)$$

We denote

$$\mathrm{var}(\underline{r}) \text{ by } \sigma^2.$$

The expected value of $\hat{\underline{b}}$ may now be obtained. We find that

$$\begin{aligned} E(\hat{\underline{b}}) &= E[(X^TX)^{-1}X^T\underline{y}] \\ &= (X^TX)^{-1}X^TE(X\underline{b} + \underline{r}) \\ &= (X^TX)^{-1}X^T(X\underline{b} + E(\underline{r})) \\ &= (X^TX)^{-1}X^TX\underline{b} \\ &= \underline{b}, \qquad (10) \end{aligned}$$

as it should be. Moreover, since

$$\begin{aligned} \hat{\underline{b}} - \underline{b} &= (X^TX)^{-1}X^T\underline{y} - \underline{b} \\ &= (X^TX)^{-1}X^T(X\underline{b} + \underline{r}) - \underline{b} \\ &= (X^TX)^{-1}X^T\underline{r}, \qquad (11) \end{aligned}$$

it follows that

$$
\begin{aligned}
\mathrm{var}(\hat{\underline{b}}) &= E[(\hat{\underline{b}} - \underline{b})(\hat{\underline{b}} - \underline{b})^T] \\
&= E[(X^TX)^{-1}X^T\underline{r}\,\underline{r}^TX(X^TX)^{-1}] \\
&= (X^TX)^{-1}X^T\ E(\underline{r}\,\underline{r}^T)\ X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}. \qquad (12)
\end{aligned}
$$

This would provide a measure of the goodness of the estimate $\hat{\underline{b}}$, if σ could be determined.

To this end we consider the vectors $\underline{y} = X\underline{b} + \underline{r}$ and $\hat{\underline{y}} = X\hat{\underline{b}} = X(X^TX)^{-1}X^T\underline{y}$ and investigate the quantity

$$S = (\underline{y} - \hat{\underline{y}})^T(\underline{y} - \hat{\underline{y}}). \qquad (13)$$

Since

$$(\underline{y} - \hat{\underline{y}}) = [I - X(X^TS)^{-1}X^T]\underline{y}, \qquad (14)$$

we have, upon noticing that $A = (I - X(X^TX)^{-1}X^T)$ is idempotent and symmetric,

$$
\begin{aligned}
S &= (A\underline{y})^TA\underline{y} \\
&= \underline{y}^TA^TA\underline{y} \\
&= \underline{y}^TA^2\underline{y} \\
&= \underline{y}^TA\underline{y} \\
&= (\underline{r}^T + \underline{b}^TX^T)[I - X(X^TX)^{-1}X^T](X\underline{b} + \underline{r}) \\
&= \underline{r}^T[I - X(X^TX)^{-1}X^T]\underline{r}. \qquad (15)
\end{aligned}
$$

The nature of S is then settled by the following theorem.

Theorem. If $\underline{x}$ is a vector of random variables with zero means and var-cov matrix σ^2I and if A is idempotent and symmetric of rank k, then

$$E(\underline{x}^TA\underline{x}) = k\sigma^2.$$

Proof:

$$E(\underline{x}^T A \underline{x}) = E\left(\sum_i x_i^2 a_{ii} + \sum_i \sum_j x_i x_j a_{ij}\right)$$

$$= \sum_i a_{ii} E(x_i^2) + \sum_i \sum_j a_{ij} E(x_i, x_j)$$

$$= \sigma^2 \sum_i a_{ii}$$

$$= \sigma^2 \, \mathrm{tr}(A)$$

$$= k\sigma^2,$$

since the trace of A is the sum of the eigenvalues, which, for an idempotent symmetric matrix, is equal to the rank. Thus $E(S) = k\sigma^2$ and

$$\sigma^2 = \frac{1}{k} E(S). \tag{16}$$

We may approximate the expected value of S by S, so that $\sigma^2 = \frac{S}{k}$. Now k is n - p, where n is the number of observations and p is the number of columns, i.e., the number of parameters. Thus the variance of $\hat{\underline{b}}$ can be estimated from known quantities.

9.3 A linear statistical model

As an illustration of the care that must be taken in formulating a model to use in analyses of the type we have been discussing, we consider the following data concerning weights of specimens of three classes of plants.

Class	Weight
Normal	101
	105
	94
Off-type	84
	88
Aberrant	32

We let y_{ij} denote the weight of the j^{th} plant in class i and assume a model

$$y_{ij} = \mu + \alpha_i + r_{ij}. \tag{17}$$

That is, we assume that the weight of a plant is scattered randomly about a constant value μ, plus a correction which depends on the class. We seek the vector $\underline{b} = [\mu, \alpha_1, \alpha_2, \alpha_3]^T$ determined by given data. For our example the system (17) has the form $\underline{y} = X\underline{b} + \underline{r}$, where

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \tag{18}$$

In detail we have:

$$\begin{aligned} \mu + \alpha_1 \qquad\qquad\quad &+ r_{11} = 101 \\ \mu + \alpha_1 \qquad\qquad\quad &+ r_{12} = 105 \\ \mu + \alpha_1 \qquad\qquad\quad &+ r_{13} = 94 \\ \mu \qquad + \alpha_2 \qquad &+ r_{21} = 84 \\ \mu \qquad + \alpha_2 \qquad &+ r_{22} = 88 \\ \mu \qquad\qquad + \alpha_3 &+ r_{31} = 32 \end{aligned} \tag{19}$$

The rank of X is three, so X^TX does not have an inverse and formula (2) is inapplicable. We can, however, still carry the analysis, as before, down to

$$(X^TS)\hat{\underline{b}} = X^T\underline{y}. \tag{20}$$

In detail, we have the inhomogeneous system

$$\begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} 504 \\ 300 \\ 172 \\ 32 \end{bmatrix}. \tag{21}$$

Gauss-Jordan elimination on the augmented matrix gives

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 32 \\ 0 & 1 & 0 & -1 & 68 \\ 0 & 0 & 1 & -1 & 54 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{22}$$

so there is an infinity of solutions

$$\hat{\underline{b}} = \begin{bmatrix} 32 \\ 68 \\ 54 \\ 0 \end{bmatrix} + k \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}. \tag{23}$$

This occurrence simply reflects the fact that μ cannot be uniquely determined in this model independent of the α_i, but rather that it is the sums $\mu + \alpha_i$ which are significant. We could have anticipated this conclusion, but, of course, the point is that in a complicated situation the mathematics will always lead to the general solution.

9.4 Notes and references

The derivation of $\underline{b} = (X^TX)^{-1}X^T\underline{y}$ is contained in most standard texts on regression theory and in several books on linear algebra, e.g., Noble [43]. The treatment of accuracy and linear models sketched here was extracted from Searle and Hausman [49] and Searle [50], where fuller accounts may be found.

LINEAR ALGEBRA

Lecture 10: Markov Chains

Mathematical topics: Matrix multiplication, eigenvalues

Comment: The concept of the Markov process is fundamental to the mathematical analysis of economic and social phenomena, as well as problems in the physical sciences. The introduction given here demonstrates the power of matrix methods in diverse applications of Markov chains.

Markov Chains

A Markov chain is a system which moves through a discrete set of states in such a way that when the system is in a state i there is a probability P_{ij} that it will next move to the state j. These possibilities are given by a transition matrix P:

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & P_{11} & P_{12} & P_{13} \\ 2 & P_{21} & P_{22} & P_{23} \\ 3 & P_{31} & P_{32} & P_{33} \end{array}$$

In this example we have chosen a 3-state system.

Example 1. The land of Oz is blessed by many things, but good weather is not one of them. The weather can be described by a Markov chain with states rain, nice, or snow, and transition matrix P:

$$\begin{array}{c|ccc} & R & N & S \\ \hline R & 1/2 & 1/4 & 1/4 \\ N & 1/2 & 0 & 1/2 \\ S & 1/4 & 1/4 & 1/2 \end{array}$$

Example 2. A study of occupational mobility in Great Britain suggests that change in occupational status from father to son could be estimated by the transition matrix P:

$$\begin{array}{c|ccc} & U & M & L \\ \hline U & .448 & .484 & .068 \\ M & .054 & .699 & .247 \\ L & .011 & .503 & .486 \end{array}$$

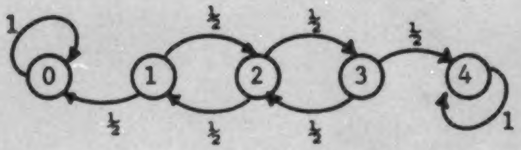
This is, it is estimated that the son of a man in the upper class has a probability .448 of remaining in the class, .484 of moving to the middle class, and .068 of moving to the lower class. A Markov chain analysis would follow a given family through several generations (assuming one in each generation and no change in the transition pattern given by P).

Example 3. (The Ehrenfest urn model for gas diffusion). We have two urns A and B which, at any moment, contain 4 balls (molecules). Each instant a ball is chosen at random and moved to the other urn. If we take as a state the number of balls in urn A, we have a Markov chain with transition matrix P:

$$\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1/4 & 0 & 3/4 & 0 & 0 \\ 2 & 0 & 1/2 & 0 & 1/2 & 0 \\ 3 & 0 & 0 & 3/4 & 0 & 1/4 \\ 4 & 0 & 0 & 0 & 1 & 0 \end{array}$$

This model was invented to describe the movement of molecules in a gas (we would have many more balls in this case). The Ehrenfests were interested in showing that in statistical equilibrium the process would appear reversible in time despite the fact that the model seems to evidence a tendency to move towards an equalization of the number of balls in each urn. Thus, for example, if one had 100 balls and were given a sequence of results, he might expect that they are in the order of increasing time if there are more pairs 57, 56 than 56, 57 and in reverse time if the contrary is true. The model shows that one would expect, in fact, equal numbers of such pairs after the system has run for some time.

Example 4. Consider a simple random walk with state 0 and four stopping states:



The transition matrix is

$$
\begin{array}{c c}
 & \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} &
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\
0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\
0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}.
$$

Consider now the first example. Suppose that today it is raining in the land of Oz. What is the probability $P_{rs}^{(2)}$ of snow two days from now? We can calculate this using the simple probability argument $P_{rs}^{(2)} = P_{rr}P_{rs} + P_{rn}P_{ns} + P_{rs}P_{ss}$. This simply takes into account all possible outcomes for the first day. But $P_{rs}^{(2)}$ is the (r,s) entry in the matrix P^2. More generally, it is easy to show for any Markov chain the probability $P_{ij}^{(n)}$ of starting in state i and ending in state j after n steps is given by the (i,j) entry of the n^th^ power P^n of P.

For example, for the weather in the land of Oz an elementary computer program in BASIC yields the powers P^2, P^4, P^8, P^{16} as follows:

```
LIST
POWER

10 READ N
20 MAT READ M(N,N)
30 FOR I = 1 TO 4
40 PRINT "POWER: "2↑I
60 MAT A = M*M
70 MAT PRINT A
80 MAT M = A
85 PRINT
90 NEXT I
95 DATA 3
100 DATA .5, .25, .25, .5, 0, .5, .25, .25, .5
200 END
READY

RUN
POWER
POWER:  2
         0.4375                     0.1875                     0.375
         0.375                      0.25                       0.375
         0.375                      0.1875                     0.4375
POWER:  4
         0.402344                   0.199219                   0.398437
         0.398437                   0.203125                   0.398437
         0.398437                   0.199219                   0.402344
POWER:  8
         0.400009                   0.199997                   0.399994
         0.399994                   0.200012                   0.399994
         0.399994                   0.199997                   0.400009
```

```
POWER: 16
        0.4                     0.2                     0.4
        0.4                     0.2                     0.4
        0.4                     0.2                     0.4
0.132 SEC.
READY
```

We notice that after 16 days the prediction of the weather becomes independent of the initial state. This is characteristic of a large class of chains, but not of all chains. For example, in the random walk problem the long-range prediction is clearly that of being in state 0 or 4, but it is intuitively clear that the chances of ending in 4 starting in 3 are much better than if we start in 1.

The behavior found in Example 1 is typical of the behavior of a regular Markov chain. This is a Markov chain such that, for some n, P^n has no zero entries. This occurs in Example 1 for $n = 2$ and in Example 2 for $n = 1$.

The principal theorem for regular chains is the following one.

Theorem. If P is regular, then $P^n \to W$, where W is a matrix all of whose rows are the same vector $\underline{\alpha}$, and $\underline{\alpha}$ is uniquely determined as a probability vector such that $\underline{\alpha}P = \underline{\alpha}$.

To find $\underline{\alpha}$ we need only solve a set of linear equations determined by $\sum_i \alpha_i = 1$, $\underline{\alpha}P = \underline{\alpha}$.

In Example 1, $\underline{\alpha} = [\frac{2}{5}, \frac{1}{5}, \frac{2}{5}]$. In Example 2, $\alpha = [.067, .624, .309]$. The theorem then shows that no matter what the initial state is, the probability of being in the upper class after a number of transitions is approximately .067, independent of the initial class. We can also interpret $\underline{\alpha}$ as follows. If the proportion in each class at any time is given by $\underline{\alpha}$, then this proportion will be maintained from that time on, since $\underline{\alpha}P^n = \underline{\alpha}$. This would be an equilibrium situation in terms of movement between the classes. At the time the study was done the actual proportions were [.076, .634, .290], indicating a slight difference from the predicted equilibrium. The model would suggest a slight downward trend from this.

A model of this sort illustrates the vast simplification made in treating mobility as a Markov chain. First we assume that in predicting the next generation only the present information is of value. Second we assume that these predictions do not change in time. The latter assumption would surely be invalidated by major social changes. Interestingly enough, the change seems to be less than one might expect, although very few studies have been made which cover more than one generation.

The proof of the basic convergence theorem for regular chains is most easily carried out by proving that if h is a column vector, then

$$P^n \underline{h} \to \underline{c},$$

where c is a column vector with constant entries. Then we must have, for example,

$$P^n \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \to \begin{bmatrix} \alpha_1 \\ \alpha_1 \\ \alpha_1 \end{bmatrix}, \qquad P^n \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \to \begin{bmatrix} \alpha_2 \\ \alpha_2 \\ \alpha_2 \end{bmatrix},$$

and

$$P^n \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \to \begin{bmatrix} \alpha_3 \\ \alpha_3 \\ \alpha_3 \end{bmatrix},$$

with the result that

$$P^n \to \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}.$$

Details of the proof may be found in the book by Noble [43]. It is a very intuitive proof which depends upon the fact that if $\underline{h}_1 = P\underline{h}$, then each component of $\underline{h}_1$ is an average of the components of $\underline{h}$ and hence the difference between the maximum and minimum value of $\underline{h}$ decreases. It actually decreases geometrically as n increases.

An alternate approach, which can be used if you have developed eigenvalue theory for nonnegative matrices, involves the normal form. For example, for three states write $P = ADA^{-1}$, where $D = \operatorname{diag}(1, \lambda_1, \lambda_2)$. The number 1 is an eigenvalue since $P\,\underline{1} = \underline{1}$ and one can show that $|\lambda| \leqq 1$. It is not always possible to obtain this simple a form, but the more complicated case causes no difficulty. Then

$$P^n = A\begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda_1^n & 0 \\ 0 & 0 & \lambda_2^n \end{bmatrix} A^{-1} \rightarrow A\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} A^{-1}$$

$$= \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix}.$$

Another quantity of interest for regular chains is

$$M_{ij} = \begin{cases} \text{the mean number of steps required to reach state } j \text{ for the first time, having started in state } i & (i \neq j) \\ 0 & (i = j) \end{cases}.$$

Let M_i = the mean time for starting in i and returning to i for the first time. Then, looking one step ahead, we obtain

$$M_{ij} = \sum_k P_{ik}(M_{kj} + 1), \text{ if } i \neq j,$$

and

$$M_i = \sum_k P_{ik}(M_{ki} + 1).$$

Using these facts, we can write the matrix equation

$$M = PM + E - D,$$

where $M = [M_{ij}]$, $E = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, and $D = \begin{bmatrix} M_1 & 0 & 0 \\ 0 & M_2 & 0 \\ 0 & 0 & M_3 \end{bmatrix}$. Multiplying by $\underline{\alpha}$ gives

$$\underline{\alpha}\, M = \underline{\alpha}\, PM + [1, 1, 1] - [\alpha_1 M_1, \alpha_2 M_2, \alpha_3 M_3].$$

Since $\underline{\alpha} P = \underline{\alpha}$,

$$[\alpha_1 M_1, \alpha_2 M_2, \alpha_3 M_3] = [1, 1, 1].$$

Hence, $M_i = \frac{1}{\alpha_i}$. Thus, in general, the mean time to return to a state is the reciprocal of the limiting probability of being in the state. For example, in the case of the land of Oz, the mean time between nice days is 5. The equations $(I - P)M = E - D$ have a unique solution. For the land of Oz example, these equations have solution

$$M = \begin{bmatrix} 0 & 4 & \frac{10}{3} \\ \frac{2}{3} & 0 & \frac{8}{3} \\ \frac{10}{3} & 4 & 0 \end{bmatrix}.$$

Absorbing Markov Chains

A state i is called absorbing if $P_{ii} = 1$. A Markov chain is called absorbing if from any state it is possible to reach an absorbing state.

Example 1. Our previous random walk example is of this type. In analyzing the behavior of an absorbing Markov chain, it is convenient to label the states in such a way that the absorbing states are first. The matrix will then have the form

$$P = \begin{bmatrix} I & 0 \\ R & Q \end{bmatrix},$$

where I is an identity matrix whose size is determined by the number of absorbing states. In our random walk example we have

$$P = \left[\begin{array}{cc:ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hdashline \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right].$$

Example 2. Consider the Leontief input-output model. Assume that we have a group of industries which produce goods for consumption by persons outside the group as well as for industries within the group. Assume that industry i requires an amount q_{ij} of industry j's goods (in dollar value) to make one dollar's worth of its goods. Let Q be the matrix $[q_{ij}]$. Then we assume (a) $Q \geqq 0$, and (b) $Q\,\underline{1} \leqq \underline{1}$. Condition (a) simply asserts that all industries have a nonnegative price for their goods, and condition (b) asserts that no industry operates at a loss. Let π be the row vector $[x_1, x_2, \ldots, x_n]$, where x_j is the output of the j^{th} industry. Let $\gamma = [c_1, c_2, \ldots, c_n]$ be the consumption vector. That is, an outside demand of c_j is made for the j^{th} industry's goods by persons outside the group of industries. To realize γ we must have

$$\pi = \pi Q + \gamma .$$

Then

$$\pi = \gamma (I - Q)^{-1} .$$

To study this we construct a Markov chain of the form

$$P = \left[\begin{array}{c|c} 1 & 0 \quad 0 \ldots 0 \\ \hline 1 - \Sigma_j \, q_{ij} & Q \end{array}\right].$$

We have simply added one ideal absorbing state. This chain will be absorbing if every industry either makes a profit or depends upon an industry that makes a profit.

Consider now the powers of an absorbing Markov chain P. We find

$$P^2 = \begin{bmatrix} I & 0 \\ R + QR & Q^2 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} I & 0 \\ R + QR + Q^2R & Q^3 \end{bmatrix}$$

$$P^n = \begin{bmatrix} I & 0 \\ (I + Q + \ldots + Q^{n-1})R & Q^n \end{bmatrix}.$$

The (i,j) entry of Q^n tends to zero because the probability of reaching an absorbing state tends to zero. Therefore $I + Q + \ldots + Q^{n-1}$ tends to $(I - Q)^{-1}$ as n increases. The (i,j) entry of $(I - Q)^{-1}$ has a simple probabilistic interpretation. It is the expected number of visits to j, starting in i (i and j nonabsorbing states). If $\underline{c}$ is the column vector with all ones, then the vector $\underline{T} = (I - Q)^{-1}\underline{c}$ gives us the mean number of steps before absorption. Finally, we see that

$$P^n \to \begin{bmatrix} I & 0 \\ (I - Q)^{-1}R & 0 \end{bmatrix},$$

so that the (i,j) entry of $(I - Q)^{-1}R$ gives the probability that the system ends in the absorbing state j, starting from the nonabsorbing state i.

The simple matrix expressions involved in these calculations are easy to program on the computer. For example, the following BASIC program gives these quantities for the random walk problem.

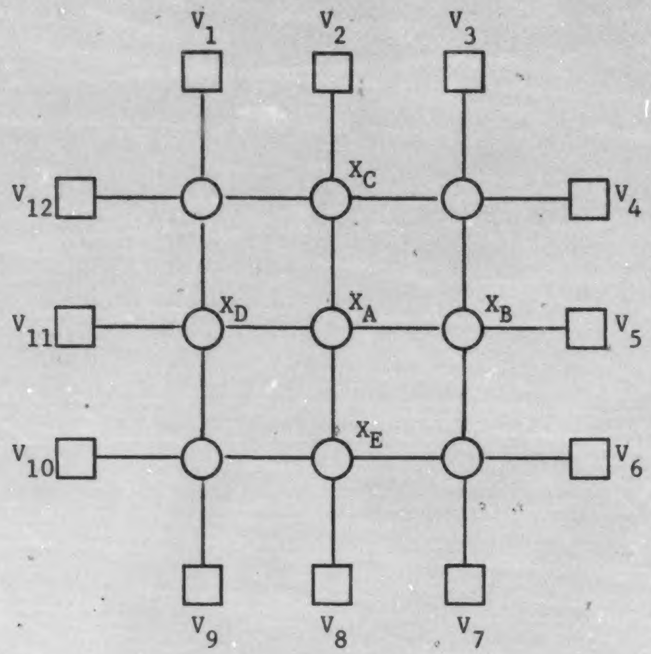
```
LIST
ACHAIN
100  READ N,M
110  REM N = NUMBER OF ABSORBING STATES, M = NUMBER OF TRANSIENT
120  MAT READ Q(M,M)
130  MAT READ R(M,N)
140  MAT C = CON(M,1)
150  MAT U = IDN(M,M)
160  MAT V = U-Q
170  MAT N = INV(V)
180  MAT T = N*C
190  MAT B = N*R
200  PRINT "TRANSIENT TO TRANSIENT PROBABILITIES"
210  MAT PRINT Q
215  PRINT
220  PRINT "TRANSIENT TO ABSORBING PROBABILITIES"
230  MAT PRINT R
235  PRINT
240  PRINT "FUNDAMENTAL MATRIX"
250  MAT PRINT N
255  PRINT
260  PRINT "ABSORPTION TIMES"
```

```
270  MAT PRINT T
275  PRINT
280  PRINT "ABSORPTION PROBABILITIES"
290  MAT PRINT B
295  PRINT
300  DATA 2,3
310  DATA 0, .5, 0, .5, 0, .5, 0, .5, 0
320  DATA .5, 0, 0, 0, 0, .5
330  END
READY


RUN
ACHAIN
TRANSIENT TO TRANSIENT PROBABILITIES
     0             0.5             0
     0.5           0               0.5
     0             0.5             0
TRANSIENT TO ABSORBING PROBABILITIES
     0.5           0
     0             0
     0             0.5
FUNDAMENTAL MATRIX
     1.5           1               0.5
     1             2               1
     0.5           1               1.5
ABSORPTION TIMES
     3
     4
     3
ABSORPTION PROBABILITIES
     0.75          0.25
     0.5           0.5
     0.25          0.75
0.163 SEC.
READY
```

We note, for example, that starting in state 1 the expected number of visits to state 2 is 1, the expected number of steps to absorption is 4, and the probability of ending up in state 4 is 3/4.

There is an interesting interplay between Markov chain theorems and electrical networks. Consider a game played on the following grid:



A player starts at a circle and moves randomly with probability $\frac{1}{4}$ to each neighboring point until he reaches a square. He then receives a payment V_k. Let X_A be the value of the game to the player when he starts at A. Then

$$X_A = \tfrac{1}{4} X_B + \tfrac{1}{4} X_C + \tfrac{1}{4} X_D + \tfrac{1}{4} X_E .$$

This means, in terms of the random walk with absorptive chain associated with the game, that

$$P\underline{x} = \underline{x}.$$

But then

$$P^n\underline{x} = \underline{x}$$

and finally

$$\begin{bmatrix} I & 0 \\ B & 0 \end{bmatrix} \underline{x} = \underline{x}.$$

If we write $\underline{x} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix}$, where $\underline{x}_1$ represents the payoffs at the absorbing states, we have

$$\underline{x}_2 = B\underline{x}_1.$$

This indicates that the value of the game at intermediate states is determined by the expected value upon reaching the absorbing states.

Assume now that the grid is an electrical network with the lines having equal resistance and the squares maintained at voltages V_k by an outside source. Let $\underline{\hat{x}}$ be the voltage vector at the various circles as well as at the squares. By Kirchhoff's law,

$$\frac{\hat{X}_B - \hat{X}_A}{R} + \frac{\hat{X}_C - \hat{X}_A}{R} + \frac{\hat{X}_D - \hat{X}_A}{R} + \frac{\hat{X}_E - \hat{X}_A}{R} = 0$$

and

$$\hat{X}_A = \tfrac{1}{4}\hat{X}_B + \tfrac{1}{4}\hat{X}_C + \tfrac{1}{4}\hat{X}_D + \tfrac{1}{4}\hat{X}_E .$$

Thus, we have two vectors, $\underline{x}$ and $\underline{\hat{x}}$, with the same average property and the same boundary values. We want to show that $\underline{x} = \underline{\hat{x}}$. To see this we first observe that a vector with the average property takes on its maximum value at the boundary. If it were to occur at an interior point, then all adjacent points would have the same value, and we could take a path to the boundary all along which the value is the same. The same argument applies to the minimum value. But $\underline{x} - \underline{\hat{x}}$ has the average property and boundary value zero. Hence $\underline{x} = \underline{\hat{x}}$.

Thus we could solve the game problem by building an electric circuit and using a voltmeter, or we could solve the electrical problem by solving the game problem. The latter can be accomplished readily with a Markov chain or by the so-called "Monte Carlo method." In the latter, one actually plays a large number of games and takes the average of the results as X_A.

The following print-out computed the interior values for a simple example by numerical methods correct to 3 decimal places in .5 seconds. The Monte Carlo method required 25 seconds with an error of as much as .006.

```
RUN
GRID
WALK OR ITERATE?  ITERATE
INPUT N?  10

      0              0              1
0     0.323          0.506          0.876        1
1     0.787          0.823          1
      1              1
TIME = 0.542496
AGAIN?  YES
WALK OR ITERATE?  WALK
INPUT N? 10000
      0              0              1
0     0.319          0.504          0.875        1
1     0.781          0.821          1
      1              1
TIME = 25.3036
AGAIN?  NO
25.324 SEC.
READY
```

Theoretical details of the discussion of this lecture can be found in Kemeny, et al. [38], Chapters 4, 5. The eigenvalue approach is found in Karlin [36].

An interesting class of problems leading from Markov chain theory deals with Markov decision theory. In this study one assumes that when the system is in state i one has a set of decisions which can be made, each decision entailing the next transition vector and also possibly an immediate reward.

One seeks to decide what to do in any state in order to maximize the total reward. An excellent reference for this problem is Howard [31].

Lecture 11: A Mechanical Vibrating System

Mathematical topics: Eigenvalues, eigenvectors, orthogonality

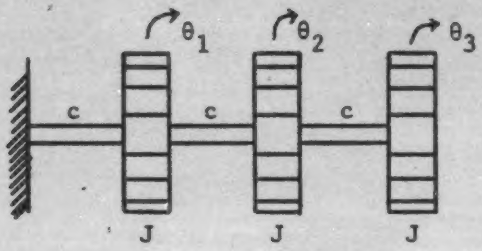
Comment: The subject of eigenvalues and eigenvectors is introduced via the study of a simple mechanical structure. The problem is treated in considerable detail as an illustration of the "case-study" approach, which would be appropriate to a class interested in this topic.

As is true of all areas of mathematics, linear algebra contains a large collection of definitions, theorems, and proofs. If these are presented in the traditional definition-theorem-proof-definition-theorem-proof- ... form, many students are soon bored and/or confused, because they are unable to discern how the definitions or theorems help to illuminate the subject. In addition, such an approach is certain to stifle interest in linear algebra in all but one or two exceptional students who are really able to understand abstractions.

In presenting linear algebra (or any other subject area of mathematics), it is helpful to study a specific problem--pure or applied--and show how, in the evolution of the solution of that problem, certain difficulties arise, and that these difficulties can be surmounted only after delving deeper into the subject. In this way one motivates the need for the theorems which must be proved.

An extremely fruitful area for such applied problems is the subject of vibrations of systems having many degrees of freedom. Here the student has a physical feeling for the subject. Everyone has seen vibrating systems, from swings in playgrounds and pendulums to long coiled springs. Some students will even know that vibratory phenomena occur in atomic physics, crystal structure, electromagnetic radiation, and in the oscillating voltages of radio and television circuits. In teaching linear algebra in a semi-concrete fashion (i.e., by choosing carefully a specific problem and carrying through the solution via linear algebra), it is very instructive to build upon the students' intuition and previous knowledge.

A problem which illustrates the above statements concerns the vibrations of three uniform disks equally spaced on a uniform shaft which is rigidly attached to a wall. Pictorially, the problem is shown below:



From the viewpoint of dynamics, the distinguishing characteristic of each disk is its polar moment of inertia J, i.e., the moment of inertia about a line perpendicular to the disk and passing through its center (assuming that each disk is a circular plate). The distinguishing characteristic of the shaft is its flexibility c, i.e., the amount of rotation that it undergoes when subjected to a moment; this is precisely the same concept for rotational motion as that of a spring constant for rectilinear motion. That is,

$$c = \frac{\text{moment needed to produce a given angular rotation}}{\text{the given angular rotation}}.$$

Usually the angular rotations are considered to be small, so that the problem remains linear (in the sense of Hooke's Law).

To solve a problem such as this--even to set up the equations of motion--using only Newton's Second Law of Motion (modified for rigid bodies, not particles; and rotational motion, not rectilinear) is very difficult. It is a formidable task to determine in the conventional manner (i.e., constructing free-body diagrams) the forces and angular accelerations acting on the individual disks.

For problems such as these (and many others), Lagrange put Newton's Second Law of Motion into a completely different form. In the absence of damping and external forces, the motion of a vibrating system--such as a simple pendulum--can be described as follows. When the pendulum hangs undisturbed from its point of support--in its equilibrium position--it has zero kinetic energy but a certain quantity of potential energy, which may be taken to be 0 because the potential energy function is determined up to a constant. After the pendulum is set in motion, at various times it arrives at its maximum excursion, when it stops moving and reverses its direction of motion. At those times the pendulum bob has zero kinetic energy, but, with respect to its equilibrium position, it has a certain quantity of potential energy. On the other hand, when the pendulum bob passes through the equilibrium position, it has zero potential energy but a certain quantity of kinetic energy. As the pendulum swings back and forth between its extreme positions, the energy is continually changed from potential to kinetic, with some of both when the pendulum bob is between the equilibrium and extreme positions. In short, in an oscillatory motion without damping and in the absence of external forces, the energy appears as purely potential or purely kinetic, or some combination of the two. Thus, in this kind of motion, the basic physical entities which are observed are not forces and accelerations (rather, moments and angular accelerations), but kinetic and potential energies. The Newton equations of motion are formulated in terms of forces and accelerations; the Lagrange equations of motion are formulated in terms of kinetic and potential energies.

In the Lagrange equations of motion applied to the problem of the three disks, the following notation is employed:

T = total kinetic energy of the system
U = total potential energy of the system
$\theta_i(t)$ = the angular displacement of disk i ($i = 1,2,3$) as a function of time t
$\dot{\theta}_i(t)$ = the angular velocity of disk i as a function of time t

Kinetic energy, loosely described as the energy due to the motion, is a function of at least the (angular) velocities of the system; and potential energy, loosely described as the energy of position, usually is a function of only the angular coordinates of the system.

The Lagrange equations of motion for an undamped system not subjected to external forces are

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta}_i}\right) + \frac{\partial U}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, n.$$

It is instructive to clarify the meanings of the separate terms in Lagrange's equations. Consider a single degree of freedom system consisting of a particle of mass m attached to a wall by means of a spring with constant K; for this system the Lagrange equation of motion reduces to

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}}\right) + \frac{\partial U}{\partial x} = 0.$$

The kinetic energy T of the particle is $\frac{1}{2}m\dot{x}^2$. Now,

$$\frac{\partial T}{\partial \dot{x}} = m\dot{x}$$

and

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}}\right) = m\ddot{x},$$

which is mass times acceleration. On the other hand, the potential energy U of the particle is $\frac{1}{2}Kx^2$, and

$$\frac{\partial U}{\partial x} = Kx,$$

which has the dimension

$$\frac{\text{force}}{\text{displacement}} \times \text{ displacement} = \text{force}.$$

Hence, dimensional analysis of Lagrange's equation of motion shows that

$$\text{force} = \text{mass} \times \text{acceleration}.$$

Thus, Lagrange's equation is another formulation of Newton's Second Law of Motion.

The great advantage of the Lagrange formulation is that in a great many problems of practical importance it is often comparatively simple to write out the expressions for the kinetic and potential energies by inspection of the diagram. For the problem of the three disks,

$$T = \tfrac{1}{2}J\left(\dot{\theta}_1^2 + \dot{\theta}_2^2 + \dot{\theta}_3^2\right)$$

$$U = \tfrac{1}{2}c\left[\theta_1^2 + (\theta_2 - \theta_1)^2 + (\theta_3 - \theta_2)^2\right].$$

(This is the place where it is sometimes desirable to motivate a discussion of the properties of quadratic forms by pointing out that both the kinetic and potential energies are examples of quadratic forms.)

With the Lagrange equations of motion it is now very easy to derive the differential equations of motion of the system:

$$J\ddot{\theta}_1 + c[\theta_1 - (\theta_2 - \theta_1)] = 0$$

$$J\ddot{\theta}_2 + c[\theta_2 - \theta_1 - (\theta_3 - \theta_2)] = 0$$

$$J\ddot{\theta}_3 + c[\theta_3 - \theta_2] = 0$$

This system of three coupled linear differential equations with constant coefficients can be written as a single matrix equation of the form

$$J \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \\ \ddot{\theta}_3 \end{bmatrix} + c \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0.$$

In proceeding from the differential equations to the matrix, several concepts can be discussed:

a) Factoring a scalar out of a matrix, i.e., the scalar is factored out of every element of the matrix--not, as in determinants, out of one row or column.

b) The square matrix is symmetric; this is a direct consequence of the fact that U is a quadratic form, which can always be written in a symmetric form.

c) The square matrix is positive-definite.

In general, it is neither necessary nor desirable to seek the most general solution of the equations of motion. Instead, it is instructive to look for special solutions, particularly for those solutions of the differential equations that represent a motion of the system during which all of the disks are oscillating with the same frequency. For this purpose, we assume that $\theta_i(t)$ is given by

$$\theta_i(t) = A_i \sin \omega t \quad (i = 1, 2, 3).$$

The verbal statement "all disks oscillate with the same frequency" is translated into an operational mathematical form by taking ω to be the same for θ_1, θ_2, and θ_3.

Setting $\theta_i(t) = A_i \sin \omega t$ into the matrix equation and canceling $\sin \omega t$ ($\neq 0$) results in

$$-J\omega^2 \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} + c \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = 0.$$

Rearranging gives

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \frac{J\omega^2}{c} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix},$$

which has the form of a typical matrix eigenvalue problem,

$$MA = \lambda A.$$

At this point it is reassuring to be able to prove that the eigenvalues $J\omega^2/c$ are positive, so that the frequencies ω will be real. Consequently, this result now becomes highly motivated, rather than simply an exercise on the part of the instructor. However, I do not know of a simple proof of this, so I prefer instead to prove the weaker result that the eigenvalues of a real symmetric matrix are real.

Theorem. The eigenvalues of a real symmetric matrix are real.

Proof: Let λ be an eigenvalue of M and let X be the associated eigenvector. By definition,

$$MX = \lambda X. \qquad (1)$$

If M is real and λ is complex, then the elements of X are also complex. Take complex conjugates of both sides of (1) to obtain

$$M\bar{X} = \bar{\lambda}\bar{X}. \qquad (2)$$

Observe that $\overline{M} = M$ because M is real. Premultiply (1) by $\overline{X}^T$ and premultiply (2) by X^T to obtain, respectively,

$$\overline{X}^T MX = \lambda \overline{X}^T X \tag{3}$$

$$X^T M\overline{X} = \overline{\lambda} X^T \overline{X}. \tag{4}$$

Now take the transpose of either (3) or (4); notice that now we motivate the question "How is the transpose of a product of matrices related to the given matrices?" and later prove that

$$(ABC)^T = C^T B^T A^T.$$

Let us take the transpose of (3), obtaining

$$X^T M\overline{X} = \lambda X^T \overline{X},$$

where we have used facts such as $M^T = M$ (because M is symmetric) and $(\overline{X}^T)^T = \overline{X}$. Subtracting (4) from (5) gives

$$0 = (\lambda - \overline{\lambda}) X^T \overline{X}.$$

Now,

$$X^T\overline{X} = [x_1, x_2, \ldots, x_n] \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \overline{x}_n \end{bmatrix}$$

$$= x_1\overline{x}_1 + x_2\overline{x}_2 + \ldots + x_n\overline{x}_n$$

$$= |x_1|^2 + |x_2|^2 + \ldots + |x_n|^2$$

$$= \sum_{r=1}^{n} |x_r|^2.$$

Since $X \neq 0$, the above expression is positive. This means that we must have $\lambda - \overline{\lambda} = 0$. Thus $\lambda = \overline{\lambda}$ and λ is real.

When the order of the matrix M is large and the eigenvalues are distinct, the determination of all of the eigenvalues is a difficult problem because the characteristic equation has a high degree (equal to the order of M). However, in many important practical problems it is not necessary to know all of the eigenvalues; sometimes the smallest or largest eigenvalue is all that is needed. In the case of vibration problems, ordinarily the smallest eigenvalue (i.e., the lowest frequency of vibration) is the most important, because if dangerous resonant vibrations are induced in a structure or system, these will occur first at the lowest frequency of vibration. For this reason it is useful to have a method of extracting from a matrix eigenvalue problem not necessarily all of the eigenvalues, but instead only a single eigenvalue--namely, the one with largest absolute value. (When the student is convinced that, in addition to being real, the eigenvalues are also positive, the modifier "absolute" can be omitted.) A procedure which is ideal for this purpose is the iterative method of determining eigenvalues, which is attractive not only because of its elegance but also because of its ease in computation.

Before it is possible to explain and derive the iterative method of determining eigenvalues, it is necessary to prove another theorem.

Theorem. The eigenvectors of a real symmetric matrix are orthogonal.

Proof: Let λ_a and λ_b be two distinct eigenvalues of a real square matrix M. Let X_a and X_b be the corresponding eigenvectors. The matrix eigenvalue problem requires the solution of

$$MX = \lambda X.$$

Then, by assumption,

$$MX_a = \lambda_a X_a$$

$$MX_b = \lambda_b X_b.$$

Premultiply the first equation by X_b^T and premultiply the second equation by X_a^T to obtain

$$X_b^T M X_a = \lambda_a X_b^T X_a$$

$$X_a^T M X_b = \lambda_b X_a^T X_b.$$

Take the transpose of both members of one of these equations--say the first one--to obtain

$$X_a^T M X_b = \lambda_a X_a^T X_b,$$

and subtract the left- and right-hand members, respectively, from the left- and right-hand members, respectively, of the second of the equations above, resulting in

$$0 = (\lambda_b - \lambda_a) X_a^T X_b.$$

By hypothesis, λ_a and λ_b are distinct eigenvalues, so it follows that

$$X_a^T X_b = 0,$$

and thus that X_a and X_b are orthogonal.

The iterative method of computing the largest eigenvalue of a real symmetric matrix M uses the fact that the eigenvectors of M are orthogonal. Because the eigenvectors are orthogonal, they form a basis for a space of n dimensions, so any vector in n-dimensional space is a linear combination of these eigenvectors.

Let the eigenvalues be ordered so that

$$\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_n.$$

Because any vector V in the space is a linear combination of the n eigenvectors $X_1, X_2, \ldots, X_n$, it follows that

$$V = \sum_{i=1}^{n} a_i X_i.$$

Then, multiplying both members of this equation by M,

$$MV = \sum_{i=1}^{n} a_i \lambda_i X_i.$$

Again multiplying by M, we obtain

$$M^2 V = \sum_{i=1}^{n} a_i \lambda_i^2 X_i.$$

Continuing m times leads to

$$M^mV = \sum_{i=1}^{n} a_i\lambda_i^m X_i$$

$$= \lambda_1^m \sum_{i=1}^{n} a_i\left(\frac{\lambda_i}{\lambda_1}\right)^m X_i$$

$$= \lambda_1^m\left[a_1X_1 + a_2\left(\frac{\lambda_2}{\lambda_1}\right)^m X_2 + \ldots + a_n\left(\frac{\lambda_n}{\lambda_1}\right)^m X_n\right].$$

If m has been chosen sufficiently large, then

$$\left(\frac{\lambda_2}{\lambda_1}\right)^m \ll 1,\quad \left(\frac{\lambda_3}{\lambda_1}\right)^m \ll \left(\frac{\lambda_2}{\lambda_1}\right)^m,\quad \ldots,\quad \left(\frac{\lambda_n}{\lambda_1}\right)^m \ll \left(\frac{\lambda_{n-1}}{\lambda_1}\right)^m,$$

and the dominant term is $a_1\lambda_1^m X_1$. Thus

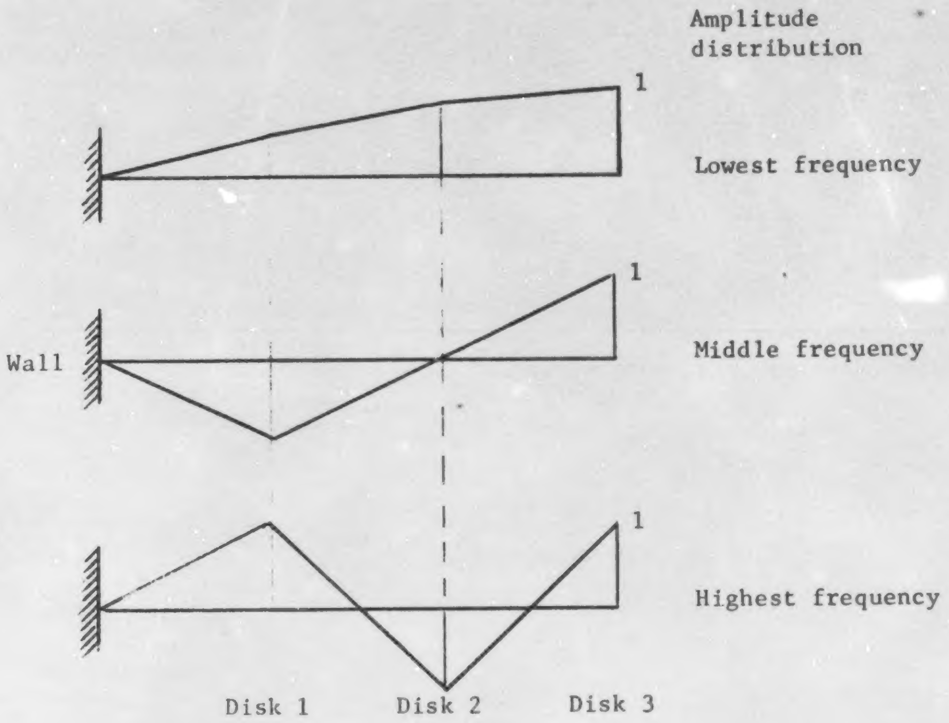
$$M^mV \cong a_1\lambda_1^m X_1.$$

The iteration method always converges to the largest eigenvalue; it also gives the eigenvector associated with the largest eigenvalue.

When choosing an initial approximation to the eigenvector associated with the largest eigenvalue, to save time and energy it is most important to make a good first guess as to the distribution of the components. For this purpose, some physical insight into the problem is very helpful. In the present problem, an elementary knowledge of vibrations is all that is needed.

Before describing a simple procedure for making a good first guess for the components of this eigenvector, it is necessary to observe that the eigenvectors are always normalized with respect to one component; i.e., all components of each eigenvector are divided by one of the components. Of course, we must be careful not to choose a component whose value may be zero. Consequently, before normalizing an eigenvector it is valuable to determine a component which can never be zero, and to use this component as the divisor. In the present problem of three disks, the disk farthest away from the wall has no resisting moment on its right and consequently will always be in motion; i.e., the amplitude A_3 of disk 3 will never be zero in any of its modes of oscillation. For this reason we choose $A_3 = 1$ in all modes of oscillation.

From physical considerations we should presume that the higher frequencies (i.e., largest eigenvalues) are associated with those processes which store the greatest amount of potential energy in the system. In general, the potential energy is largest when the twist in the shafts is greatest, i.e., when the disks attempt to oscillate in opposite directions, in this way twisting the shafts as much as possible. Building up from the lowest frequencies, it is then evident that a sketch of the three modes of oscillation of the system will be as follows.



From the sketches we can see that a reasonable first approximation X_{3_o} to the components of the eigenvector X_3 associated with the largest eigenvalue λ_3 (i.e., the highest frequency of oscillation) is

$$X_{3_o} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

When this first approximation to X_3 is set into the matrix eigenvalue problem, there results a first approximation X_{3_1}

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix} = 2 \begin{bmatrix} 3/2 \\ -4/2 \\ 2/2 \end{bmatrix} ;$$

here, 2 is the first approximation to λ_3 and $\begin{bmatrix} 15 \\ -2 \\ 1 \end{bmatrix}$ is the first approximation to X_3. Now discard the 2 (after recording it, just to keep track of the approximations to λ_3 to determine whether they converge), and in $MX = \lambda X$ replace the zero approximation X_{3_o} by the first approximation

$$X_{3_1} = \begin{bmatrix} 3/2 \\ -4/2 \\ 2/2 \end{bmatrix} ,$$

which is left in fractional form to facilitate computation. This results in

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3/2 \\ -4/2 \\ 2/2 \end{bmatrix} = \begin{bmatrix} 10/2 \\ -13/2 \\ 6/2 \end{bmatrix} = 3 \begin{bmatrix} 10/6 \\ -13/6 \\ 6/6 \end{bmatrix} ,$$

so that 3 is the second approximation to λ_3 and

$$\begin{bmatrix} 10/6 \\ -13/6 \\ 6/6 \end{bmatrix}$$

is the second approximation to X_3. From $\lambda_{3_o} = 2$ and $\lambda_{3_1} = 3$ it is clear that the iterative procedure has not yet converged, so it is necessary to continue the process. Continuing in this way to obtain X_{3_2}, multiply X_{3_1} by M to obtain

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 10/6 \\ -13/6 \\ 6/6 \end{bmatrix} = \begin{bmatrix} 33/6 \\ -42/6 \\ 19/6 \end{bmatrix} = \frac{19}{6} \begin{bmatrix} 33/19 \\ -42/19 \\ 19/19 \end{bmatrix}.$$

Then $19/6 \cong 3.166$ is a third approximation to λ_3 and the above vector is a third approximation to X_3. After this procedure is carried out several more times, the sequence of eigenvalues $\lambda_{3_1}, \lambda_{3_2}, \ldots, \lambda_{3_m}$ approaches 3.245, so that the highest frequency ω_3^2 is

$$\omega_3^2 = 3.245 \text{ c/J}.$$

The eigenvector corresponding to ω_3^2 is

$$A^{(3)} = \begin{bmatrix} 1.800 \\ -2.242 \\ 1.000 \end{bmatrix}.$$

In a problem such as this one, it is not difficult to use the iterative procedure just outlined to compute the smallest eigenvalue and the corresponding eigenvector. For this purpose, it is necessary to obtain a matrix eigenvalue problem of the form

$$MA = \frac{1}{\lambda} A;$$

the determination of the largest eigenvalue $\frac{1}{\lambda}$ of this matrix eigenvalue problem is equivalent to obtaining the smallest eigenvalue λ.

In the present problem, when we seek to determine the lowest frequency of the system, we want to compute, instead of $J\omega^2/c$, the value of $c/J\omega^2$. There are several ways of determining a matrix eigenvalue problem which contains $c/J\omega^2$ instead of $J\omega^2/c$. One procedure is simply to premultiply both sides of

$$MA = \lambda A$$

by the inverse of M and obtain

$$A = \lambda M^{-1}A;$$

divide by λ and end with the matrix eigenvalue problem

$$M^{-1}A = \lambda^{-1}A.$$

This merely uses the fact that if λ is an eigenvalue of A and A has an inverse, then λ^{-1} is an eigenvalue of A^{-1}. This procedure is awkward because it requires determination of the inverse of M.

Another method of setting up the matrix eigenvalue problem which has for its largest eigenvalue a number proportional to the lowest frequency of oscillation involves use of the so-called "influence coefficients." Because the angular rotations are small, they are (by Hooke's Law) linearly proportional to the moments. If a unit moment is applied at disk 1, that disk will rotate through a unit angle; because no moment is applied at disks 2 and 3, the portion of the shaft to the right of disk 1 will rotate as a rigid body and hence will also rotate through a unit angle. This is put into mathematical form by writing a matrix of influence coefficients which has each entry of its first row equal to 1. If a moment of 2 units is applied at disk 2, that disk will rotate through two unit angles, but disk 1, being halfway between the wall and disk 2, will rotate only half as much, i.e., through 1 unit angle, and disk 3 will rotate through the same angle as disk 2. This means that the second row of the matrix of influence coefficients is (1,2,2). Finally, if a moment of 3 units is applied at disk 3, because of the linear nature of the angular deflection as a function of the moment, disk 1 will rotate through 1 unit angle, disk 2 through 2 unit angles, and disk 3 through 3 unit angles. Hence, the third and final row of the matrix of influence coefficients is (1,2,3). Thus, the matrix of influence coefficients has the form

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix};$$

it can be shown without too much effort that this is precisely the inverse of

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Then, multiplying

$$J \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \\ \ddot{\theta}_3 \end{bmatrix} + c \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0$$

by

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

results in

$$J \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \\ \ddot{\theta}_3 \end{bmatrix} + c \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0.$$

Again assuming that we seek the normal modes of oscillation, we set

$$\theta_i(t) = A_i \sin \omega t, \quad i = 1, 2, 3,$$

and, after cancellation of the common term $\sin \omega t$ ($\neq 0$), we obtain

$$-J\omega^2 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} + c \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = 0.$$

This in turn reduces to the matrix eigenvalue problem

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \frac{c}{J\omega^2} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} ,$$

exactly the form desired.

Another way of accomplishing the same result is to rewrite the equations of motion in the form

$$\begin{aligned} 2\theta_1 - \theta_2 \qquad &= -J\ddot{\theta}_1/c \\ -\theta_1 + 2\theta_2 - \theta_3 &= -J\ddot{\theta}_2/c \\ -\theta_2 + \theta_3 &= -J\ddot{\theta}_3/c \end{aligned}$$

and solve for θ_1, θ_2, and θ_3 in terms of $\ddot{\theta}_1$, $\ddot{\theta}_2$, and $\ddot{\theta}_3$, resulting in

$$\begin{aligned} \theta_1 &= -\frac{J}{c}(\ddot{\theta}_1 + \ddot{\theta}_2 + \ddot{\theta}_3) \\ \theta_2 &= -\frac{J}{c}(\ddot{\theta}_1 + 2\ddot{\theta}_2 + 2\ddot{\theta}_3) \\ \theta_3 &= -\frac{J}{c}(\ddot{\theta}_1 + 2\ddot{\theta}_2 + 3\ddot{\theta}_3) \ . \end{aligned}$$

The matrix form of these three linear differential equations is

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = -\frac{J}{c} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \\ \ddot{\theta}_3 \end{bmatrix} .$$

Now set

$$\theta_i(t) = A_i \sin \omega t, \quad i = 1, 2, 3,$$

and divide both sides by $J\omega^2/c$ to obtain

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \frac{c}{J\omega^2} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix},$$

which is identical with the result obtained above.

To determine the largest eigenvalue of this problem (i.e., the smallest value of ω) by iteration, once more we choose an initial approximation to the eigenvector corresponding to this eigenvalue. For this purpose, we refer back to the figure giving amplitude distribution and choose as an initial approximation

$$A_{1_o} = \begin{bmatrix} 0.5 \\ 0.8 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 5/10 \\ 8/10 \\ 10/10 \end{bmatrix}.$$

Again, it should be noted that the normalization is on the third component of A_1, i.e., on the angular rotation of disk 3.

Starting with A_{1_o}, the first approximation to A_1 is

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 5/10 \\ 8/10 \\ 10/10 \end{bmatrix} = \begin{bmatrix} 23/10 \\ 41/10 \\ 51/10 \end{bmatrix} = \frac{51}{10} \begin{bmatrix} 23/51 \\ 41/51 \\ 51/51 \end{bmatrix} \cong 5.1 \begin{bmatrix} 0.451 \\ 0.805 \\ 1.000 \end{bmatrix}.$$

The second approximation to A_1 is obtained by discarding the 5.1 (which now is the first approximation to the smallest eigenvalue λ_1) and multiplying the column vector by the matrix of influence coefficients to give

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 23/51 \\ 41/51 \\ 51/51 \end{bmatrix} = \begin{bmatrix} 115/51 \\ 207/51 \\ 258/51 \end{bmatrix} = \frac{258}{51} \begin{bmatrix} 115/258 \\ 207/258 \\ 258/258 \end{bmatrix} \cong 5.05 \begin{bmatrix} 0.445 \\ 0.806 \\ 1.000 \end{bmatrix}.$$

Because the initial guess was such a good one, the iteration process has essentially produced results accurate to three significant figures after two iterations. If the procedure is continued another time, the result is the third approximation to A_1, or

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 115/258 \\ 207/258 \\ 258/258 \end{bmatrix} = \begin{bmatrix} 580/258 \\ 1045/258 \\ 1303/258 \end{bmatrix} = \frac{1303}{258} \begin{bmatrix} 580/1303 \\ 1045/1303 \\ 1303/1303 \end{bmatrix} \cong 5.06 \begin{bmatrix} 0.445 \\ 0.801 \\ 1.000 \end{bmatrix}.$$

Only one more iteration is needed to obtain a result accurate to three significant figures:

$$5.06 \begin{bmatrix} 0.445 \\ 0.802 \\ 1.000 \end{bmatrix}$$

Thus the largest value of $c/J\omega^2$ is 5.06,

$$\omega_1^2 = 0.198\ c/J,$$

and the distribution of amplitudes in the lowest mode is

$$\begin{bmatrix} 0.445 \\ 0.802 \\ 1.000 \end{bmatrix}.$$

Now it is necessary to obtain an eigenvalue intermediate between the smallest and largest, which have already been obtained. In a relatively simple problem such as this one involving only three disks (i.e., 3 degrees of freedom) leading to a matrix eigenvalue problem where the matrix is only 3 x 3, it is not difficult to obtain the characteristic equation (which is of degree 3) and factor out the linear factors involving the smallest and largest eigenvalues. This leads to a linear equation which is easy to solve.

In the present problem, the characteristic equation is

$$\begin{vmatrix} 1 - \lambda & 1 & 1 \\ 1 & 2 - \lambda & 2 \\ 1 & 2 & 3 - \lambda \end{vmatrix} = -(\lambda^3 - 6\lambda^2 + 5\lambda - 1) = 0,$$

corresponding to the problem where the eigenvalues in order of increasing size ($\omega_1 < \omega_2 < \omega_3$) are $c/J\omega_1^2$, $c/J\omega_2^2$, and $c/J\omega_3^2$. Corresponding to the problem where the eigenvalues $J\omega_3^2/c$, $J\omega_2^2/c$, and $J\omega_1^2/c$ are given in order of increasing size, the characteristic equation is

$$\begin{vmatrix} 2 - \mu & -1 & 0 \\ -1 & 2 - \mu & -1 \\ 0 & -1 & 1 - \mu \end{vmatrix} = -(\mu^3 - 5\mu^2 + 6\mu - 1) = 0.$$

One of these characteristic equations can be obtained from the other by the simple transformation $\lambda = 1/\mu$.

Because the iteration method of computing eigenvalues always converges to the largest eigenvalue, this procedure cannot be used to compute intermediate eigenvalues. This means that, except under highly improbable circumstances, any initially chosen eigenvector which is assumed to approximate the distribution of amplitudes in a mode other than the highest will converge to the eigenvector associated with the largest eigenvalue. The only exception is the circumstance that, by an extremely lucky guess, the problem-solver happens to choose, as a first approximation to the distribution of amplitudes in the lower mode, precisely that distribution of amplitudes. In this event, the iteration procedure "locks onto" the initial guess and cannot produce any new distributions of amplitudes, i.e., the first guess is already a solution, and the iterative procedure does not converge to the eigenvector associated with the largest eigenvalue. Instead, it is necessary to modify the iteration method by introducing additional information about the eigenvectors.

The additional information needed is that the eigenvectors of a real symmetric matrix are orthogonal. In general form, it has already been observed that any vector V in the space of dimension equal to the order of the square matrix can be written as a linear combination of the eigenvectors of the matrix, i.e.,

$$V = \sum_{i=1}^{n} a_i A_i .$$

Here a_1 is the coefficient of the eigenvector associated with the largest eigenvalue of the matrix, a_2 is the coefficient of the eigenvector associated with the second largest eigenvalue of the matrix, etc. Then, to eliminate the highest mode of oscillation (i.e., the eigenvector associated with the largest eigenvalue), we must force a_1 to be 0, in this way removing from the problem the eigenvector associated with the largest eigenvalue.

To this end, we note that in the highest mode of oscillation the distribution of amplitudes (i.e., the components of A_3, the eigenvector associated with the largest eigenvalue) is known. For an oscillating system with n degrees of freedom, we denote these amplitudes by

$$A_1^{(n)}, A_2^{(n)}, \ldots, A_n^{(n)}.$$

Also denote the distribution of amplitudes in the second highest mode of oscillation by

$$A_1^{(n-1)}, A_2^{(n-1)}, \ldots, A_n^{(n-1)}.$$

Because of the orthogonality of the eigenvectors,

$$A_1^{(n)}A_1^{(n-1)} + A_2^{(n)}A_2^{(n-1)} + \ldots + A_n^{(n)}A_n^{(n-1)} = 0.$$

(Recall: $A_1^{(n)}, A_2^{(n)}, \ldots, A_n^{(n)}$ is the distribution of amplitudes of the highest mode; these have been determined. $A_1^{(n-1)}, A_2^{(n-1)}, \ldots, A_n^{(n-1)}$ is the distribution of amplitudes of the second-highest mode; these are to be determined.)

The above equation can be solved for one of the unknowns

$$A_1^{(n-1)}, A_2^{(n-1)}, \ldots, A_n^{(n-1)};$$

to be specific, we solve for $A_1^{(n-1)}$ and obtain

$$A_1^{(n-1)} = -\frac{A_2^{(n)}A_2^{(n-1)}}{A_1^{(n)}} - \frac{A_3^{(n)}A_3^{(n-1)}}{A_1^{(n)}} - \cdots - \frac{A_n^{(n)}A_n^{(n-1)}}{A_1^{(n)}} .$$

We now set up a matrix equation

$$\begin{bmatrix} A_1^{(n-1)} \\ A_2^{(n-1)} \\ A_3^{(n-1)} \\ \vdots \\ A_n^{(n-1)} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{A_2^{(n)}}{A_1^{(n)}} & -\frac{A_3^{(n)}}{A_1^{(n)}} & \cdots & -\frac{A_n^{(n)}}{A_1^{(n)}} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} A_1^{(n-1)} \\ A_2^{(n-1)} \\ A_3^{(n-1)} \\ \vdots \\ A_n^{(n-1)} \end{bmatrix} ,$$

which can be written in the form

$$A^{(n-1)} = SA^{(n-1)}.$$

The square matrix S is called the "sweeping matrix," because it "sweeps out" the highest mode of oscillation of the system, and the second largest eigenvalue of the original problem then becomes the largest eigenvalue of the modified problem and now can be obtained by iteration.

In the problem at hand, with $A_1^{(2)}$, $A_2^{(2)}$, and $A_3^{(2)}$ as the amplitude distribution in the second lowest mode, the orthogonality condition requires that we employ the components (0.445, 0.802, 1.000) of the eigenvector associated with the lowest frequency of oscillation. Thus,

$$0.445\ A_1^{(2)} + 0.802\ A_2^{(2)} + 1.000\ A_3^{(2)} = 0$$

and

$$A_1^{(2)} = -1.800\ A_2^{(2)} - 2.247\ A_3^{(2)}.$$

In matrix form we have

$$\begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & -1.800 & -2.247 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} .$$

Alternatively, starting with the components (1.800, -2.242, 1.000) of the eigenvector associated with the highest frequency of oscillation, the orthogonality condition leads to the equation

$$A_1^{(2)} = -1.245\ A_2^{(2)} - 0.555\ A_3^{(2)}$$

and to the matrix equation

$$\begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & -1.245 & -0.555 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} .$$

Either procedure leads to the same result.

Now we return to the matrix form of the equations of motion, i.e.,

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \lambda \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} ,$$

where $\lambda = c/J\omega^2$ (this equation shows that we are seeking the distribution of amplitudes in the second mode). We replace the column matrix on the left by the sweeping matrix times the same column matrix to obtain

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 & -1.800 & -2.247 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \lambda \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} ,$$

which leads to

$$\begin{bmatrix} 0 & -0.800 & -1.247 \\ 0 & 0.200 & -0.247 \\ 0 & 0.200 & 0.753 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \lambda \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} .$$

Now choose an initial distribution of amplitudes corresponding to a reasonable approximation to the expected second mode. Because $A_1^{(2)}$ is always multiplied by 0, there is no need to choose an initial $A_1^{(2)}$; we choose

$$A_o^{(2)} = \begin{bmatrix} — \\ 0 \\ 1 \end{bmatrix}$$

and substitute this into the above expression. We obtain

$$\begin{bmatrix} 0 & -0.800 & -1.247 \\ 0 & 0.200 & -0.247 \\ 0 & 0.200 & 0.753 \end{bmatrix} \begin{bmatrix} — \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.247 \\ -0.247 \\ 0.753 \end{bmatrix} = 0.753 \begin{bmatrix} -1.655 \\ -0.328 \\ 1.000 \end{bmatrix} .$$

Applying the matrix iteration procedure only a few times gives

$$\lambda_2 = c/J\omega_2^2 = 0.6433$$

$$A_1^{(2)} = -1.247, \ A_2^{(2)} = -0.555, \ A_3^{(2)} = 1.000.$$

The sweeping matrix can have other forms. One is obtained by solving the orthogonality condition equation for $A_3^{(2)}$ rather than $A_1^{(2)}$, resulting in

$$A_3^{(2)} = -0.445 \ A_1^{(2)} - 0.802 \ A_2^{(2)} .$$

The matrix form then becomes

$$\begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -0.445 & -0.802 & 0 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} .$$

Setting this sweeping matrix into the matrix form of the equations of motion results in

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -0.445 & -0.802 & 0 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0.555 & 0.198 & 0 \\ 0.110 & 0.396 & 0 \\ -0.335 & -0.406 & 0 \end{bmatrix} \begin{bmatrix} A_1^{(2)} \\ A_2^{(2)} \\ A_3^{(2)} \end{bmatrix} .$$

Now when we choose an initial approximation to the eigenvector A_2, we first observe that $A_3^{(2)}$ is always multiplied by 0. On the basis of the sketch of the expected second mode, we make the initial guess

$$A_1^{(2)} = -1, \; A_2^{(2)} = 0, \; A_3^{(2)} = — .$$

Setting this into the matrix equation, we obtain

$$\begin{bmatrix} 0.555 & 0.198 & 0 \\ 0.110 & 0.396 & 0 \\ -0.335 & -0.406 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \\ — \end{bmatrix} = \begin{bmatrix} -0.555 \\ -0.110 \\ 0.335 \end{bmatrix} = 0.335 \begin{bmatrix} -1.654 \\ -0.328 \\ 1.000 \end{bmatrix} .$$

Continuing with the normalized amplitudes $A_1^{(2)} = -1.654$, $A_2^{(2)} = -0.328$, $A_3^{(2)} = -$, there results

$$\begin{bmatrix} 0.555 & 0.198 & 0 \\ 0.110 & 0.396 & 0 \\ -0.335 & -0.406 & 0 \end{bmatrix} \begin{bmatrix} -1.654 \\ -0.328 \\ — \end{bmatrix} = \begin{bmatrix} -0.987 \\ -0.312 \\ 0.683 \end{bmatrix} = 0.683 \begin{bmatrix} -1.445 \\ -0.457 \\ 1.000 \end{bmatrix} .$$

After this process has been repeated several times, the results coincide with those obtained by use of the previous sweeping matrix; i.e., the process converges to the same eigenvalue (frequency of oscillation) and amplitude distribution (normalized eigenvector).

If the smallest eigenvalue were not already known, it could be determined by application of two orthogonality conditions. For this purpose we make use of the previous results.

Largest eigenvector: $A_1^{(1)} = 0.445$, $A_2^{(1)} = 0.802$, $A_3^{(1)} = 1.000$.

Second largest eigenvector: $A_1^{(2)} = -1.247$, $A_2^{(2)} = -0.555$, $A_3^{(2)} = 1.000$.

With $A_1^{(3)}$, $A_2^{(3)}$, and $A_3^{(3)}$ as the components of the eigenvector associated with the lowest mode of vibration, we now obtain two orthogonality conditions:

$$0.445\, A_1^{(3)} + 0.802\, A_2^{(3)} + 1.000\, A_3^{(3)} = 0$$

$$-1.247\, A_1^{(3)} - 0.555\, A_2^{(3)} + 1.000\, A_3^{(3)} = 0$$

These orthogonality conditions can be combined to give

$$1.6923\, A_1^{(3)} + 1.3572\, A_2^{(3)} = 0,$$

so that

$$A_1^{(3)} = 1.800\, A_3^{(3)},\quad A_2^{(3)} = -2.245\, A_3^{(3)}.$$

This can be put into the form of a new sweeping matrix, which, when incorporated into the equations of motion, serves to "sweep out" both the lowest and second lowest modes of oscillation:

$$\begin{bmatrix} A_1^{(3)} \\ A_2^{(3)} \\ A_3^{(3)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1.800 \\ 0 & 0 & -2.245 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_1^{(3)} \\ A_2^{(3)} \\ A_3^{(3)} \end{bmatrix}$$

This is set into

$$\begin{bmatrix} 0.555 & 0.198 & 0 \\ 0.110 & 0.396 & 0 \\ -0.335 & -0.406 & 0 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \lambda \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix},$$

to obtain

$$\begin{bmatrix} 0.555 & 0.198 & 0 \\ 0.110 & 0.396 & 0 \\ -0.335 & -0.406 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1.800 \\ 0 & 0 & -2.245 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0.555 \\ 0 & 0 & -0.691 \\ 0 & 0 & 0.308 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix}.$$

Notice that in the last matrix equation both $A_1^{(3)}$ and $A_2^{(3)}$ are multiplied by 0, so guesses as to their initial values are not needed; therefore, we choose an initial distribution of amplitudes

$$A_1^{(3)} = -,\ A_2^{(3)} = -,\ A_3^{(3)} = 1.000.$$

After one step of the iteration procedure we obtain $c/J\omega_3^2 = 0.308$, so that

$$\omega_3^2 = 3.245\ c/J,\ A_1^{(3)} = 1.800,\ A_2^{(3)} = -2.242,\ A_3^{(3)} = 1.000.$$

Thus, the final results are:

Lowest frequency	Middle frequency	Highest frequency
$\omega_1^2 = 0.198\ c/J$	$\omega_2^2 = 1.555\ c/J$	$\omega_3^2 = 3.245\ c/J$
$A^{(1)} = \begin{bmatrix} 0.445 \\ 0.802 \\ 1.000 \end{bmatrix}$	$A^{(2)} = \begin{bmatrix} -1.247 \\ -0.555 \\ 1.000 \end{bmatrix}$	$A^{(3)} = \begin{bmatrix} 1.800 \\ -2.242 \\ 1.000 \end{bmatrix}$

There are several checks that can be made in these results. For instance, it is possible to assure oneself that the eigenvalues are feasible, using two theorems proved in Gantmacher [20]:

Theorem. The sum of the eigenvalues of a square matrix equals the trace of the matrix. (The trace of a matrix is the sum of the elements on the principal diagonal.)

Theorem. The product of the eigenvalues of a square matrix equals the determinant of the matrix.

For the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

we have that the trace of A equals 6 and the determinant of A is 1. With

$$\lambda_1 = 0.198,\ \lambda_2 = 1.555,\ \lambda_3 = 3.245,$$

we have

$$\sum_{i=1}^{3} \lambda_i = 0.198 + 1.555 + 3.245 = 4.998$$

$$\lambda_1\lambda_2\lambda_3 = (0.198)(1.555)(3.245) \cong 1.000,$$

results which are sufficiently accurate to ensure confidence in the computation of the eigenvalues.

We can also check the orthogonality of the eigenvectors

$$A^{(1)} = \begin{bmatrix} 0.445 \\ 0.802 \\ 1.000 \end{bmatrix},\quad A^{(2)} = \begin{bmatrix} -1.247 \\ -0.555 \\ 1.000 \end{bmatrix},\quad A^{(3)} = \begin{bmatrix} 1.800 \\ -2.242 \\ 1.000 \end{bmatrix},$$

obtaining

$$\sum_{i=1}^{3} A_i^{(1)}A_i^{(2)} = (0.445)(-1.247) + (0.802)(-0.555) + (1)(1) \cong 0$$

$$\sum_{i=1}^{3} A_i^{(1)}A_i^{(3)} = (0.445)(1.800) + (0.802)(-2.242) + (1)(1) \cong 0.001$$

$$\sum_{i=1}^{3} A_i^{(2)}A_i^{(3)} = (-1.247)(1.800) + (-0.555)(-2.242) + (1)(1) \cong -0.001.$$

Again, this check is sufficiently accurate to ensure confidence in the computation of the eigenvectors.

Lecture 12: Algebraic Coding Theory

Suppose we want to transmit a message consisting of a string of k letters, each letter being 0 or 1. Further suppose that our transmission lines are noisy, so that some of the 1's may be received as 0's and vice versa. We would like a rule whereby we could add r so-called check digits to the message which would enable the receiver to determine with reasonable assurance what the original message was. Note that there are only 2^k possible messages but 2^n possible words to be sent, where $n = k + r$. We define the information rate R to be

$$R = \frac{k}{n}.$$

We would like to maximize R while at the same time keeping the probability of a decoding error to a minimum.

Example 1. Suppose $k = 1$ and we decide to add 2 check digits according to the rule of repetition, i.e., if the message is 1 we send the word 1 1 1 and if the message is 0 we send 0 0 0. Let us assume that the transmission lines make an error (i.e., change a 1 to a 0 or a 0 to a 1) with probability 1/10. Then if the word received is 1 1 1 and we guessed that the message was a 1, we would be correct with probability .999. If the word received was 1 1 0 or 1 0 1 or 0 1 1, we should again guess that the message sent was a 1 and we would be correct with probability .900. We could decrease the chance for an error by adding more and more check digits, but this would be done at the expense of the information rate, which is $\frac{1}{r+1}$.

Example 2. This is called a parity check code. Suppose we have a message of length k and we wish to add a single check digit. We choose the check digit so as to make the total number of 1's in the word be even. For example, if we want the message to be 1 0 1 we send the word 1 0 1 0, and if we want to send the message 1 0 0 we send out 1 0 0 1. In this case our information rate, $\frac{k}{k+1}$, is high, but if there is an even number of errors there will be a decoding failure. Even if you could detect an error, you would not know whether there were more than one, and even if you assumed that there was only one error, you would not know where it occurred.

Linear algebra enters the picture when we think of the word being sent as an n-tuple of 0's and 1's and the transmission error as an n-tuple of 0's and 1's, where a 1 in the i^{th} place means that an error was made in that place. So if

$$\underline{C} = [C_1, \ldots, C_n]$$

is the word sent, and

$$\underline{E} = [E_1, \ldots, E_n]$$

is the transmission error, then the received word will be

$$\underline{R} = [R_1, \ldots, R_n]$$

with

$$R_i = C_i + E_i,$$

where addition is understood to be modulo 2. (Thus our words are to be thought of as vectors in the n-dimensional space over the field $Z_2 = \{0,1\}$.)

Suppose we wish k to be 3 and r to be 3, i.e., we wish C_1, C_2, C_3 to be our message digits and C_4, C_5, C_6 to be our check digits. Suppose we choose our check digits according to the rules

$$C_4 = C_1 + C_2$$

$$C_5 = C_1 + C_3$$

$$C_6 = C_2 + C_3,$$

where addition is again modulo 2. We can rewrite these equations as

$$C_1 + C_2 + C_4 = 0$$

$$C_1 + C_3 + C_5 = 0$$

$$C_2 + C_3 + C_6 = 0$$

or, in matrix form, as

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \underline{C}^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

or

$$H\underline{C}^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \qquad (1)$$

The matrix H is called the parity check matrix, and a code obtained in this way in called a linear code. Note that the number of rows of H is r, and the number of columns of H is n. Also note that we may consider H to be a linear map from the vector space of n-tuples to the vector space of r-tuples (where all the entries are from a field with 2 elements).

Observe that the rows of the parity matrix are linearly independent, so the map must be onto. Equation (1) and the fact that $\underline{R} = \underline{C} + \underline{E}$ gives us

$$H\underline{R}^T = H\underline{E}^T.$$

The receiver has knowledge of H and so can construct $H\underline{R}^T$ and hence $H\underline{E}^T$. This vector is called the syndrome. His task is to construct $\underline{E}$ and thus form $\underline{C} = \underline{R} + \underline{E}$.

Now, unfortunately, two error vectors may give rise to the same syndrome (the map $X \rightarrow HX$ is not one-to-one). However, since the map is linear, once we have any one possible error vector, all the other possible ones are obtained from the first one by adding the kernel of H (which, we note, consists of all the codewords). We then choose the error vector with the least number of 1's, again assuming that the probability for an error is slight.

In our example the eight codewords are

000000, 001011, 010101, 011110,

1C1101, 110011, 111000, 100110.

If the syndrome is 011, then we see immediately that one possible error is 000011. The other possible errors are obtained by adding to this error all of the codewords. Thus we get as possible error vectors

000011, 001000, 010110, 011101,
101110, 110000, 111011, 100101,

of which we choose 001000 as the most likely error. We therefore change the third digit of the received word to get what we think was the correct message. (Mathematically speaking, all the error vectors giving rise to a fixed syndrome form a coset and we want to pick a coset representative of least weight.)

Exercise 1. What is the parity check matrix of example 1?

Exercise 2. What is the parity check matrix of example 2?

We will now construct a linear code which will correct single errors. This code is called a Hamming code after its discoverer.

We note that if H is a parity check matrix with columns $r_1, \ldots, r_n$ and if $\underline{S}$ is the syndrome arising from the error $\underline{E} = [E_1, \ldots, E_n]$, then

$$\underline{S} = \sum_{i=1}^{M} E_i r_i .$$

Thus, if any column of the matrix is $\underline{0}$, then an error in that position will have no effect on that syndrome. If two columns of the parity check matrix are identical, then a single error in one of those two positions will yield the same syndrome as a single error in the other. Both single error patterns lie in the same coset. Since only one of them may be chosen as coset representative, we will have what is called a decoding failure. Thus a linear code is incapable of correcting all patterns of single errors unless all the columns of its H matrix are distinct and nonzero. Conversely, if all the columns are distinct and nonzero, single errors in different positions result in different syndromes. Note that we may read off the position of the error just by seeing which column arises as the syndrome. (Of course if two errors occur, we would have a decoding failure).

Suppose we wish to have r check digits. Then clearly the maximum number of distinct nonzero columns is $2^r - 1$. Hence the information rate is

$$R = \frac{2^r - 1 - r}{2^r - 1} .$$

The following is a Hamming code matrix for $r = 3$:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Note that the columns are the numbers 1 through 7 in binary form (reading up).

We want to find linear codes which will correct all double error patterns. That is, we will assume that at most two errors occur in our transmission and we wish to find the positions where they occur by looking at the syndrome. Again let us take $M = 2^r - 1$, but this time it turns out that we need 2r check digits, so that

$$R = \frac{2^r - 1 - 2r}{2^r - 1} .$$

Let us take $r = 3$ for definiteness. In the course of deriving this code it will be necessary to add, multiply, and divide 3-tuples according to the laws of ordinary arithmetic (with the exception that $(1, 0, 0) + (1, 0, 0) = (0, 0, 0)$, where $(1, 0, 0)$ is our unit of multiplication). Thus we are dealing with a field of 8 elements. It is an algebraic fact that there exist fields of 2^r elements for any r, and the elements may be written out concretely and addition and multiplication tables given. Usually the elements are represented as polynomials of degree $r - 1$ with coefficients in Z_2. Addition is as usual and multiplication is modulo some fixed irreducible polynomial of degree r.

We would like to have a double-error correcting code matrix that looks like

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ j_1(1) & j_1(2) & & \cdots & & j_1(7) \\ j_2(1) & j_2(2) & & \cdots & & j_2(7) \\ j_3(1) & j_3(2) & & \cdots & & j_3(7) \end{bmatrix}.$$

Viewing the matter in this way, we seek a function j which maps 3-tuples onto 3-tuples.

Let us try functions j for which the last 3 rows are also a Hamming matrix. (If one Hamming matrix corrects one error, perhaps 2 Hamming matrices will correct 2 errors.) Thus, we want j to be a permutation.

Let us suppose that our two errors occur in positions β_1 and β_2. Considering β_1 and β_2 as 3-tuples, we will get a syndrome

$$\begin{bmatrix} T_1 \\ T_2 \end{bmatrix},$$

where T_1 and T_2 are two 3-tuples written vertically to give a 6-tuple; then

$$\beta_1 + \beta_2 = T_1$$

$$j(\beta_1) + j(\beta_2) = T_2.$$

The decoder's problem is to determine β_1 and β_2, given the above set of equations. (He knows, of course, what the syndrome is.)

Let us now consider several possible choices for j. If $j(\beta) = \alpha\beta$ for some fixed 3-tuple α, the second equation just becomes a multiple of the first, and nothing new is added by the second equation. Similarly for $j(\beta) = \beta + \alpha$.

Next we try $j(\beta) = \beta^2$. The two equations for the decoder are then

$$\beta_1 + \beta_2 = T_1 \tag{1}$$

$$\beta_1^2 + \beta_2^2 = T_2. \tag{2}$$

But in characteristic 2, the left side of (1) is just the square of the left side of (2), and so $T_2 = T_1^2$. Again the second set of parity equations gives us no new information.

Now try $j(\beta) = \beta^3$. The decoder's equations then become

$$\beta_1 + \beta_2 = T_1$$

and

$$\beta_1^3 + \beta_2^3 = T_2.$$

Cubing each side of the first equation and manipulating, we get

$$\begin{aligned} \beta_1 + \beta_2 &= T_1 \\ \beta_1\beta_2 &= T_1^2 + \frac{T_2}{T_1}, \quad \text{if } T_1 \neq 0. \end{aligned}$$

Hence β_1 and β_2 are the roots of the quadratic equation

$$\beta^2 + T_1\beta + (T_1^2 + \frac{T_2}{T_1}) = 0,$$

which can be solved.

But T_1 is never equal to zero unless there were no errors and the syndrome is $\underline{0}$. For if there were two errors, then $\beta_1 \neq \beta_2$ and so $T_1 = \beta_1 + \beta_2 \neq 0$. And if there were one error in the β_1 position, then $T_1 = \beta_1$ and $T_2 = \beta_1^3 = T_1^3$. (If there were two errors we could not have $T_2 = T_1^3$ because then our quadratic equation would become

$$\beta^2 + T_1\beta + \left(T_1^2 + \frac{T_2}{T_1}\right) = \beta^2 + T_1\beta + \left(T_1^2 + \frac{T_1^3}{T_1}\right)$$

$$= \beta^2 + T_1\beta = 0.$$

But this has only one nonzero solution, namely $\beta = T_1$.)

Reference: Berlekamp [6].

LINEAR ALGEBRA

Lecture 13: Relativistic Quantum Mechanics

The key mathematical idea of the relativistic theory of the electron, as developed by P. A. M. Dirac in 1928, illustrates some of the similarities and differences between matrices and numbers. It is easily understood by sophomore students of linear algebra.

Quantum mechanics may be formulated in such a way that differential operators correspond to physically observable quantities. The fundamental operator required is the one corresponding to the energy, usually denoted H. The basic law is expressed in the famous Schrödinger equation

$$H \psi(\underline{x},t) = \frac{\partial}{\partial t} \psi(\underline{x},t). \tag{1}$$

Physical reasoning consistent with the theory of relativity led to the following expression for the square of the energy of a free particle:

$$H^2/c^2 = \nabla^2 + (mc^2)^2. \tag{2}$$

Here c is the speed of light, m is the mass of the particle, and ∇^2 is the Laplacian $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. To take the square root of this expression is not apparently meaningful. Relativity imposes the additional constraint that this root must be linear in partial derivatives, which would seem only to make the impossibility more evident.

Dirac, however, proceeded to ask that all this be done. Specifically, he wrote

$$H/c = \alpha_1 \frac{\partial}{\partial x} + \alpha_2 \frac{\partial}{\partial y} + \alpha_3 \frac{\partial}{\partial z} + \alpha_4 \; (mc^2) \tag{3}$$

and demanded that (3), when squared, should give (2). The latter will occur if the coefficients satisfy

$$\alpha_i^2 = 1 \quad (i = 1,2,3,4)$$

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 0 \quad (i, j = 1, \ldots, 4, \quad i \neq j) \qquad (4)$$

as direct computation shows. Dirac realized that equations (4), though impossible for numerical coefficients, can be readily fulfilled by matrices. A solution of smallest order is

$$\alpha_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \qquad \alpha_2 = \begin{bmatrix} 0 & 0 & 0 & i \\ 0 & 0 & -i & 0 \\ 0 & i & 0 & 0 \\ -i & 0 & 0 & 0 \end{bmatrix}$$

$$\alpha_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \qquad \alpha_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} .$$

Equation (1) thus becomes a 4 × 4 matrix differential equation or, equivalently, a system of partial differential equations.

A matrix interpretation of the Schrödinger equation had not been previously contemplated, and the fact that $\underline{\psi}$ has to be a four-component vector might have dissuaded most investigators from taking seriously the matrix solution. In the hands of Dirac, however, the four-component nature of the solution first yielded a theoretical explanation of so-called electron spin, which explains the magnetic properties of matter and is fundamental to understanding the periodic table of elements, and, second, led Dirac to predict the existence of anti-matter some seven years before the first anti-particle was discovered.

LINEAR ALGEBRA

Lecture 14: Artificial Inbreeding

Example 1. Suppose that it is desirable on a farm to have a certain type of plants as nearly alike as possible by obtaining plants which are similar homozygotes, that is, all AA or all aa. Let us suppose that the plants are subjected to continual self-fertilization. Let a_o, b_o, and c_o be the proportions of the original quantity of plants which are of types AA, Aa, and aa, respectively. Note that $a_o + b_o + c_o = 1$. The AA plants will produce only AA plants and the aa plants yield only aa plants. However, of the plants produced by Aa plants, ¼ will be AA, ½ will be Aa, and ¼ will be aa. If a_1, b_1, and c_1 represent the proportions of AA, Aa, and aa, respectively, in the first generation produced, then

$$X_1 = PX_o,$$

where

$$X_o = \begin{bmatrix} a_o \\ b_o \\ c_o \end{bmatrix}, \quad X_1 = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix}, \quad P = \begin{bmatrix} 1 & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix}.$$

Similarly,

$$X_2 = PX_1 = P^2X_o.$$

In general,

$$X_n = P^nX_o, \quad n = 1, 2, 3, \ldots.$$

If we let

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix},$$

then

$$B^{-1} = \begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

Let

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then

$$X_n = B^{-1}D^nBX_o,$$

where

$$D^n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2^{-n} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

so that

$$P^n = \begin{bmatrix} 1 & 2^{-1} - 2^{-(n+1)} & 0 \\ 0 & 2^{-n} & 0 \\ 0 & 2^{-1} - 2^{-(n+1)} & 1 \end{bmatrix}.$$

In equation form we have

$$a_n = a_o + (2^{-1} - 2^{-(n+1)})b_o$$

$$b_n = 2^{-n}b_o$$

$$c_n = c_o + (2^{-1} - 2^{-(n+1)})b_o.$$

As n approaches ∞ we have

$$a_n \to a_0 + \tfrac{1}{2} b_0$$

$$b_n \to 0$$

$$c_n \to c_0 + \tfrac{1}{2} b_0.$$

Reference: Smith [51].

Example 2. Suppose we now select a plant of type AA and use it to fertilize all of the plants as well as succeeding generations of plants. In this case we have

$$X_n = P^n X_0, \quad n = 1, 2, 3, \ldots,$$

where

$$P = \begin{bmatrix} 1 & \tfrac{1}{2} & 0 \\ 0 & \tfrac{1}{2} & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Now

$$P = B^{-1}DB,$$

where

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \tfrac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Let B = I + T, where

$$T = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then

$$\begin{aligned} B^{-1} &= (I + T)^{-1} \\ &= I - T + T^2 - T^3 + \dots \\ &= I - T + T^2 \quad \text{(since } T^n = 0 \text{ for } n \geqq 3) \\ &= \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Thus

$$P^n = \begin{bmatrix} 1 & 1 - 2^{-n} & 1 - 2^{-n+1} \\ 0 & 2^{-n} & 2^{-n+1} \\ 0 & 0 & 0 \end{bmatrix}.$$

In equation form we have

$$a_n = a_o + (1 - 2^{-n})b_o + (1 - 2^{-n+1})c_o$$

$$b_n = 2^{-n}b_o + 2^{-n+1}c_o$$

$$c_o = 0.$$

As n approaches ∞,

$$a_n \to a_o + b_o + c_o$$

$$b_n \to 0$$

$$c_n \to 0.$$

If the fertilizing plant is chosen to be of the form aa, then we need only interchange the a_i's and c_i's, $i = 0, 1, 2, \ldots$, to obtain a solution.

LINEAR ALGEBRA

Bibliography

This is not a systematic bibliography of books that deal with applications of linear algebra at the sophomore and junior level. Rather it is a selection from among the books which we have encountered in preparing these notes that could serve as sources of further applications.

1. Allen, Roy G. *Mathematical Economics*, 2nd ed. New York, St. Martin's Press, Inc., 1959.

   A classic reference which includes applications of matrix theory in connection with the theory of games, linear programming, allocation of resources, the theory of the firm.

2. Almon, Clopper. *Matrix Methods in Economics*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1967.

   Covers the mathematics of regression and linear and nonlinear programming in compact fashion with little reference to economics. Discusses computer program for regression.

3. Amundsen, Neal R. *Mathematical Methods in Chemical Engineering: Matrices and their Applications*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1966.

4. Bartholomew, D. J. *Stochastic Models for Social Processes*. New York, John Wiley and Sons, Inc., 1968.

   Discusses models for social mobility.

5. Bellman, Richard. *Stability Theory of Differential Equations*. New York, Dover Publications, 1969.

6. Berlekamp, Elwyn R. *Algebraic Coding Theory*. New York, McGraw-Hill Book Company, 1968.

   A specialist book, but the first chapter involves mainly simple matrix algebra.

7. Bickley, William G. and Thompson, R. S. *Matrices: Their Meaning and Manipulation*. New York, Van Nostrand Reinhold Company, 1964.

   The first chapter gives a number of simple applications of matrices to electrical and mechanical systems.

8. Boot, Johannes. *Mathematical Reasoning in Economics and Management Science*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1967.

Contains a variety of simple applications including Leontief models and Markov chains.

9. Braae, R. *Matrix Algebra for Electrical Engineers*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1964.

Clear and simple exposition.

10. Brauer, Fred, *et al*. *Linear Mathematics*. Menlo Park, California, W. A. Benjamin, Inc., 1970.

An excellent example of the recent trend to teach linear algebra and illustrate it by applying the theory to systems of ordinary differential equations.

11. Brouwer, W. *Matrix Methods in Optical Instrument Design*. Menlo Park, California, W. A. Benjamin, Inc., 1964.

An unusual application of matrices to ray-tracing.

12. Crowell, R. H. and Williamson, R. E. *Calculus of Vector Functions*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1962.

Calculus of several variables in the spirit of linear algebra.

13. Dantzig, George B. *Linear Programming and Extensions*. Princeton, New Jersey, Princeton University Press, 1963.

A basic reference on linear programming with many applications.

14. Davies, Owen L., ed. *Design and Analysis of Industrial Experiments*, 2nd ed. New York, Hafner Publishing Company, 1967.

A chemical engineering optimization application involving quadratic forms is given in the last chapter by G. E. P. Box.

15. Eisenschitz, Robert K. *Matrix Algebra for Physicists*. New York, Plenum Publishing Corporation, 1966.

Gives several applications of matrices to physics, including atomic oscillations and statistical mechanics.

16. Feeman, George F. and Grabois, Neil R. *Linear Algebra and Multivariable Calculus*. New York, McGraw-Hill Book Company, 1970.

An attempt to integrate linear algebra and multivariate calculus.

17. Forsythe, George E. and Moler, C. B. Computer Solution of Linear Algebraic Systems. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1967.

An elementary and authoritative treatment.

18. Frazer, Robert A., et al. Elementary Matrices and Some Applications to Dynamics and Differential Equations. New York, Cambridge University Press, 1938.

A classic, but applications are somewhat technical.

19. Gale, David. Theory of Linear Economic Models. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1961.

Excellent treatment of linear programming and Leontief models at an advanced level.

20. Gantmacher, Felix R. Theory of Matrices, vol. I. New York, Chelsea Publishing Company, 1960.

21. Gass, Saul I. Linear Programming. New York, McGraw-Hill Book Company, 1969.

Contains one chapter of applications.

22. Glicksman, Abraham M. Linear Programming and the Theory of Games. New York, John Wiley and Sons, Inc., 1963.

An interesting, simple, yet nontrivial, treatment.

23. Goldberg, Samuel. Introduction to Difference Equations. New York, John Wiley and Sons, Inc., 1958.

The last chapter is devoted to applications of matrix methods.

24. Goldstein, Herbert. Classical Mechanics. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1950.

Contains standard material on small vibrations, moments of inertia, and change of coordinates.

25. Goodman, A. W. and Ratti, J. S. Finite Mathematics with Applications. New York, The Macmillan Company, 1971.

Clear and elementary. The theory is interlined with social science applications.

26. Graybill, Franklin A. Introduction to Linear Statistical Models, vol. 1. New York, McGraw-Hill Book Company, 1961.

Uses matrices systematically throughout.

203

27. Graybill, Franklin A. *Introduction to Matrices with Applications in Statistics*. Belmont, California, Duxbury Press, 1969.

Clear treatment of some mathematical topics useful in statistics, not usually found in standard texts.

28. Guillemin, Ernst A. *Mathematics of Circuit Analysis*. Cambridge, Massachusetts, M.I.T. Press, 1949.

Matrices applied to electrical networks.

29. Heading, J. *Matrix Theory for Physicists*. New York, John Wiley and Sons, Inc., 1958.

Applications in geometry and mechanics.

30. Higman, Bryan. *Applied Group-Theoretic and Matrix Methods*. New York, Dover Publications, 1964.

Contains some interesting applications, mainly to theoretical physics and chemistry.

31. Howard, Ronald A. *Dynamic Programming and Markov Processes*. Cambridge, Massachusetts, M.I.T. Press, 1960.

32. Hyvarinen, L. P. *Mathematical Modeling for Industrial Processes*. New York, Springer-Verlag New York, 1970.

Applications are written out longhand but can be converted into matrix notation.

33. Intriligator, Michael D. *Mathematical Optimization and Economic Theory*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1971.

Good reference text at a more advanced level.

34. Johnston, J. B., *et al*. *Linear Equations and Matrices*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1966.

Two good chapters on applications in social science, business, and economics.

35. Johnston, John. *Econometric Methods*. New York, McGraw-Hill Book Company, 1963.

Regression and linear models, using matrices systematically throughout.

36. Karlin, S. *Mathematical Methods and Theory in Games, Programming and Economics*. vols. I and II. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1959.

Advanced, but excellent.

37. Kemeny, John G., *et al*. *Finite Mathematics with Business Applications*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1962.

A variant of [38], with business applications.

38. Kemeny, John G., *et al*. *Introduction to Finite Mathematics*, 2nd ed. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1966.

Chapters on linear programming, theory of games, and Markov chains.

39. Kemeny, John G. and Snell, J. Laurie. *Mathematical Models in the Social Sciences*. Waltham, Massachusetts, Blaisdell Publishing Company, 1962.

A source of interesting examples.

40. Keyfitz, Nathan. *Introduction to the Mathematics of Population*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1968.

Matrices used to find age-distribution of population subsequent to census.

41. Manne, Alan S. *Economic Analysis for Business Decisions*. New York, McGraw-Hill Book Company, 1961.

Contains a variety of specific examples involving linear programming.

42. Noble, B. *Applications of Undergraduate Mathematics in Engineering*. New York, The Macmillan Company, 1967.

43. Noble, Ben. *Applied Linear Algebra*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1969.

44. Nodelman, H. M. and Smith, F. W. *Mathematics for Electronics with Applications*. New York, McGraw-Hill Book Company, 1956.

Elementary treatment of ladder networks and transistors.

45. Pestel, Edward C. and Leckie, Frederick A. *Matrix Methods in Elastromechanics*. New York, McGraw-Hill Book Company, 1963.

Advanced and authoritative.

46. Pielou, E. C. *Introduction to Mathematical Ecology*. New York, John Wiley and Sons, Inc., 1969.

Matrices used throughout.

47. Pipes, Louis A. *Matrix Methods for Engineering*. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1963.

A miscellaneous collection of examples.

48. Rogers, Andrei. Matrix Analysis of Interregional Population Growth and Distribution. Berkeley, California, University of California Press, 1968.

Straightforward applications of matrix multiplication.

49. Searle, S. R. and Hausman, W. H. Matrix Algebra for Business and Economics. New York, John Wiley and Sons, Inc., 1970.

Applications to regression and linear statistical models.

50. Searle, Shayle R. Matrix Algebra for the Biological Sciences. New York, John Wiley and Sons, Inc., 1966.

Applications to regression and linear statistical models.

51. Smith, C. A. Biomathematics: The Principles of Mathematics for Students of Biological and General Science, 2nd ed. New York, Hafner Publishing Company, 1966.

52. Yan, C. S. Introduction to Input-Output Economics. New York, Holt, Rinehart and Winston, Inc., 1969.

Clear and elementary introduction with applications.

DIFFERENTIAL EQUATIONS

Lecture 1: General Lecture

We have come together to explore applications of elementary differential equations, in the hope that the experience will enrich our teaching of this subject. We shall begin with very simple equations and gradually add complexity. We shall try to bring the mathematics and applications together so closely that each aspect helps us to better understand the other: by seeing a mathematical problem physically, we can anticipate properties of the solutions; by recognizing a physical problem as a special case of a known mathematical problem, we see at once what to expect in the physical situation. Thus above all we should be able to unify many physical and mathematical ideas.

This morning I shall introduce the subject by pointing out some general ideas and surveying what I hope we can cover in our eight days together. In general, I am counting on you to do much of the work. I have brought along materials which we can look at and then make individual assignments.

Differential equations appear in the sciences because the derivative or integral is a valuable model for certain aspects of the real world. We should recognize at the start that we are speaking only of a model, that is of a precise mathematical idea which in a rough way imitates the physical phenomenon. For example, we speak of velocity of a physical object and represent it by a derivative, say dx/dt. However, the precise physical interpretation involves us with subtle questions of frames of reference, measurement of time, relativity, what we mean by a physical object, and so on. On a microscopic scale, an object such as a stone is a collection of molecules, and the velocity of the stone can only have meaning as some sort of average. On an even finer scale, we are dealing with atoms, and then nuclear particles, whose identity as particles is blurred with that of waves by modern quantum mechanics. Physics has provided no final answer to this question, and the physicists themselves continually shift from one model to another. At the extreme is the model of classical mechanics which suffices for many of the everyday applications. In this model we allow for bona fide particles, mass-points, and rigid bodies, and assume we have a frame of reference in which velocities can be measured.

The same difficulties appear in social sciences, where dx/dt might be the rate of change of the cost of living. The world of economics is usually understood as a discrete one--one is dealing with transactions in terms of multiples of a unit--of dollars or units of goods, of land, etc. Thus again we have invented a new model in which we use the derivative and think of dx/dt as the derivative of a real-valued function. Similarly, dx/dt might refer to the rate of change of a population, again measured in units.

We could give many, many examples of the same type of problem. In essentially all applications of the calculus to the real world, we are dealing with a convenient, but not fully precise, description of the real world. The precision can be so good as to give the illusion of perfection, as in astronomy where planetary motion is predicted with fantastic accuracy. However, absolute perfection is never claimed nor is it needed.

With this background in mind, we return to differential equations. We mentioned the importance of the derivative. It appears throughout physics in describing motion: velocity, acceleration; also in describing the rate of change of such quantities as density, temperature, energy, electric field strength, linear and angular momentum; further in describing the variation in space of the same quantities--here we are dealing with gradients, directional derivatives, Laplacians, etc. It appears in other physical sciences in similar ways--for example, in chemistry in describing reaction rates or rates of change of mass with respect to time. The derivative is also used in the biological sciences to describe time variations in the nervous system, for example. It appears in economics as stated above, also in other social sciences such as psychology--for example, to measure rate of learning.

While we have emphasized the derivative, we could equally well have used the integral $\int_{t_0}^{t_1} x(t)\,dt$, for example. It would be easy to recast most applications completely in terms of the integral. Historically, this has generally not been done, and we shall not do it here. However, mathematically there are real advantages in doing so. Integration is easier to analyze mathematically, and computational procedures usually relate to integration rather than differentiation.

An ordinary differential equation is an equation

$$F(x, y, y', \ldots, y^{(n)}) = 0 \qquad (1)$$

relating an unspecified real function y(x) and its derivatives up to a certain order n --the order of the differential equation. We shall consider only equations solvable for the highest derivative:

$$y^{(n)} = g(x, y, y', \ldots, y^{(n-1)}). \tag{2}$$

The equation can then be written as a system:

$$\begin{aligned} &\frac{dy_1}{dx} = y_2, \quad \frac{dy_2}{dx} = y_3, \ldots, \frac{dy_{n-1}}{dx} = y_n, \\ &\frac{dy_n}{dx} = g(x, y_1, y_2, \ldots, y_n), \end{aligned} \tag{3}$$

where $y_1(x) = y(x)$, $y_2(x) = y'(x)$, ..., $y_n(x) = y^{(n-1)}(x)$.
More generally one can consider a system:

$$\frac{dy_i}{dx} = F_i(x, y_1, \ldots, y_n) \quad (i = 1, \ldots, n), \tag{4}$$

or, in vector notation,

$$\frac{d\vec{y}}{dx} = F(x,\vec{y}). \tag{5}$$

All the problems we shall consider deal with such differential equations.

Our purpose here is not to go into the basic mathematical theory of ordinary differential equations. In teaching the subject at an elementary level, this theory is far too difficult. However, some of the results of the theory should be stated without proof. One goal of the elementary course should then be to prove these results for special cases.

A solution of an equation (2) is a function $y(x)$ defined over an interval which satisfies the equation identically over the interval:

$$y^{(n)}(x) = g\Big(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\Big), \; a < x < b.$$

There is a similar definition for (4) or (5), where a solution is a vector function $\big(y_1(x), \ldots, y_n(x)\big)$. The mathematical theory then tells us that, under reasonable hypotheses on the functions g or F_i, there is a unique solution satisfying given initial conditions: for (4) the initial conditions have the form $y(x_o), \ldots, y^{(n-1)}(x_o)$; for (5) they have the form $\big(y_1(x_o), \ldots, y_n(x_o)\big)$. We refer to this theorem as "the existence theorem" and do not pursue it further here. However, we shall have many occasions to illustrate it, and the theorem (but not its proof) should receive much stress in an elementary course.

In many of the applications, the independent variable is time, and we then use t rather than x. We shall change notation freely--and students should also be flexible on this.

The existence theorem itself has much significance for applications. First of all, it shows us that a given differential equation, say (2), describes a class of functions. It is thus reasonable that a "law of nature" should be described by a differential equation, for such a law normally describes restrictions on how nature may function, without fully specifying the exact behavior; it describes a class of allowed behaviors--hence (via models) a class of functions. For example, the law for falling bodies near the earth, $m\frac{d^2x}{dt^2} = - mg$, describes all motions of a particular particle in falling.

Secondly, the existence theorem contains an idea of causality. If the independent variable is time t, then the solution of the equation, say $x(t)$, is determined completely by its values and derivatives, up to a certain order, at an initial time t_o. For the falling body, $x(t)$ is uniquely determined for all t by $x(t_o)$ and the velocity $x'(t_o)$. Similarly, for a system of particles in classical mechanics the motion of the system for all time is completely determined by the positions and velocity vectors of all particles at time t_o. This is illustrated by the motion of the planets in the solar system. One is led from this result to questions of determinacy and free will. The existence theorem does not fully answer these questions since, as stressed before, we are dealing with a crude model for the real world. In fact, other models do allow for less determinism; in physics, some models are statistical and predict only average behavior. Others, as in quantum mechanics, say that precise knowledge of both position and velocity is impossible. This is not the sort of question to treat exhaustively in an elementary course on differential equations. However, it should be mentioned, along with a statement that the question is far from settled by physicists or philosophers. Einstein, in particular, could never fully accept modern quantum mechanics, which he helped to found, because of the uncertainties included in the theory.

The existence theorem contains another aspect, which is a similar source of difficulty of a philosophical nature. For the case of the planets--or any similar system of particles--the theorem implies that if we could reverse all velocity vectors at a given time, then the future motion would be just the past motion retraced backwards. For the falling body it is easy to verify this experimentally. However, for a large system of particles this "reversibility" of the motion appears to conflict with observed statistical features of the motion (i.e., statistical mechanics)--the "entropy" should always increase. Again, the question is still unsettled, and very probably it has to do with the model chosen.

The existence theorem points out that solutions of a differential equation exist. The next question is how we find the solutions. In this era of the computer, there is a very simple answer: each particular solution can be computed to desired accuracy by a numerical process. Even in an elementary course some time should be spent on the numerical procedures and, if computer facilities are available, in illustrating them on the computer. It is also instructive to do some simple problems by the Euler method of "step-by-step" integration. For example,

$$\frac{dx}{dt} = x^2 + t^2, \quad x(0) = 1,$$

could be solved with, say, $\Delta t = 0.5$, from $\Delta x \sim (x^2 + t^2)\Delta t$:

t	x	x^2+t^2	Δx
0.0	1.0	1.0	0.5
0.5	1.5	2.5	1.25
0.1	2.75		etc.

[For a more extensive discussion of this technique see, for example, Feynman [11].]

This is easily extended to systems and to n^{th}-order equations and is of much value in justifying the existence theorem. It should be pointed out that, in using such a numerical method for an applied problem, one is in effect changing the model, replacing a continuous one by a discrete one.

It is very tempting to let the computer take over completely at this point and to say that, for all practical purposes, all ordinary differential equations have been solved and the computer will read off the solutions you request. We are, in fact, essentially in possession of a vast table of solutions to all ordinary differential equations. If we could imagine this table filling a library--volume after volume, page after page of figures--we would see that our understanding of the solutions is still very incomplete. The numbers alone, by their abundance, would confuse us rather than enlighten us.

What is desperately needed is some way to organize the knowledge and for us to relate it to the applications which gave rise to the differential equations in question.

The process of organizing the knowledge is carried out, not by examining the numerical solutions (except for a few special cases), but by studying the differential equations as part of mathematics and by establishing properties which are of importance: periodicity, boundedness, stability (of many kinds), resonance. It is these properties which we shall emphasize. We shall consider only a few special types of equations and their applications:

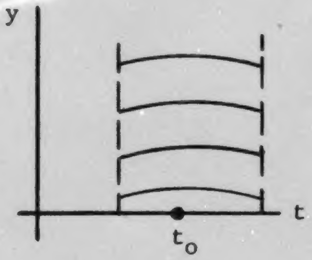
1. $y' = f(t)$
2. $y' = ky$
3. $y' = f(y) \quad \left(y' = f(y)g(t)\right)$
4. $y' + ay = f(t)$
5. $y'' + ay' + by = f(t)$
6. $y_i' = \Sigma a_{ij}y_i + f_i(t)$
7. Special nonlinear differential equations

The applications will come from many fields, especially physics and electrical engineering.

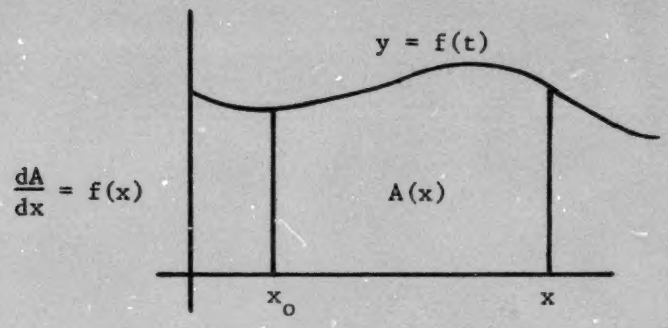
The equation $y' = f(t)$. This is the simplest differential equation, and it is important to recognize it as such. The calculus provides an existence theorem (for continuous f):

$$y(t) = \int_{t_o}^{t} f(u)\,du + g(t_o).$$

The solutions are parallel curves.



The applications are the familiar ones encountered in calculus: area,



volume, mass, moments, etc. One can lump together many applications by pointing out that one is studying processes of accumulation--of mass, of money, of population. By traveling, one is accumulating distance. One should encourage students to make simple experiments such as the following. When traveling in a car, note the speed every five minutes, and then estimate the total distance traveled in, say, one hour. This leads to a numerical integration of the speed $f(t)$. Similarly, one can observe the rate of flow of traffic past a certain point at various times (at 7 a.m., 9 a.m., 11 a.m., etc.) and integrate to estimate the total number of vehicles passing the point in a day. Waiting in line at a cafeteria, one can estimate how fast the line moves (feet per minute) before reaching the food section and how fast it moves at different points in that section. There are many other practical examples, besides those found in calculus books.

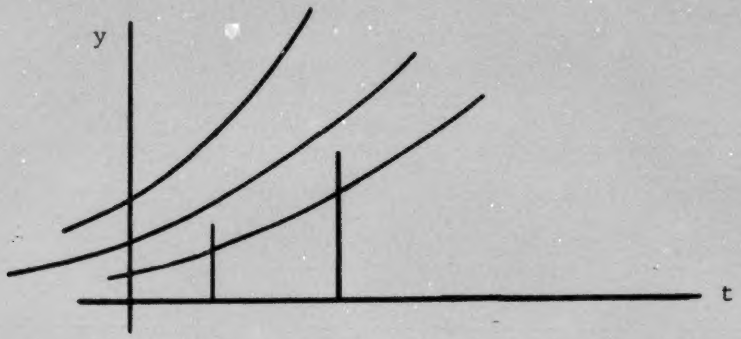
The equation $y' = ky$. This equation occurs in a variety of growth and decay problems. One should point out that the equation is unaffected by a change of y-units: $u = ay$ satisfies $u' = ku$, and the only equation $y' = f(y)$ with this property is $y' = ky$, for we want $f(u) = \frac{1}{a} f(au)$ for all $a > 0$; hence $f(u) = u\, f(1) = ku$.

It is instructive to carry out a limit process for various applications. For populations, imagine that at time t there are y persons (or animals or bacteria). At time $t + \Delta t$, Δt small, the normal process of births would add a certain number, clearly roughly proportional to y and to Δt, i.e., $\Delta y \sim ky\, \Delta t$; for a rough limit we get $y' = ky$. This could be illustrated with numbers. The process of death is similar, only now we have

$\Delta y \sim k_1 y\, \Delta t - k_2 y\, \Delta t$ [where the first term $k_1 y\, \Delta t$ represents births and the second term $-k_2 y\, \Delta t$ represents deaths]; thus $y' = (k_1 - k_2)y$. One should illustrate by numbers the fact that good accuracy is obtained even though y is an integer--the process is really statistical. The solutions of the equation are

$$y = y_o e^{kt},$$

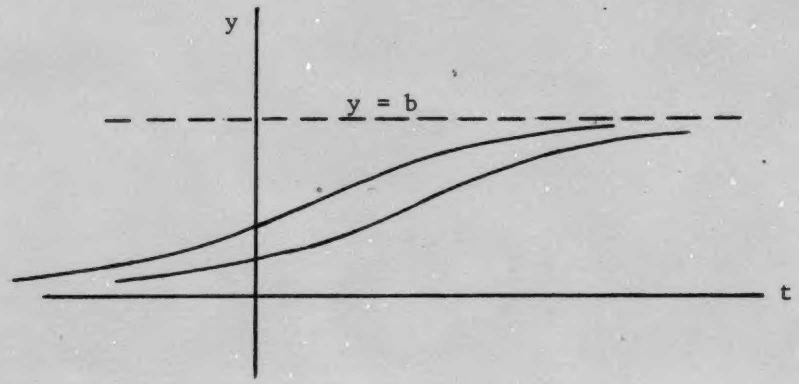
where $k = k_1 - k_2$. These solutions are unbounded as t approaches $+\infty$; hence they must be qualified for real populations.



Notice that all of the solution curves can be obtained from one piece of one curve by changing the y-scale.

Other equations of the type $y' = f(y)$. If the population is restricted by the environment, then there is a maximum number, b, which can be supported. One is then led to an equation

$$\frac{dy}{dt} = ky(b - t), \quad 0 < y < b.$$



A similar discussion applies to the "Law of Mass Action" in chemistry which leads to the differential equation

$$\frac{dz}{dt} = k(A - z)\ (B - z),\ 0 \leqq z < A < B.$$

See Golomb and Shanks [14].

DIFFERENTIAL EQUATIONS

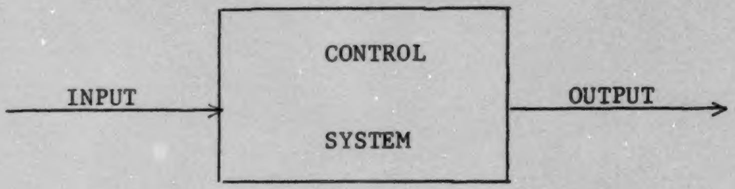
Lecture 2: Feedback Control Systems

References:

Kaplan [17], Sections 3-1, 3-2, 3-3, 3-6, 3-7, 3-9, 3-10, 3-11.

Dorf [10], Chapter 1.

The general ideas of feedback control systems first became important during World War II when engineers began to construct new control devices by putting together existing integrators. The fundamental ideas are: input, control system, output.



We will stress the engineering applications, but we will also consider applications in economics and biology.

Example. Suppose we try to drive a car at a uniform speed (input). We observe that the actual speed (output) does not conform exactly to the desired speed.

We can interpret a first-order linear differential equation

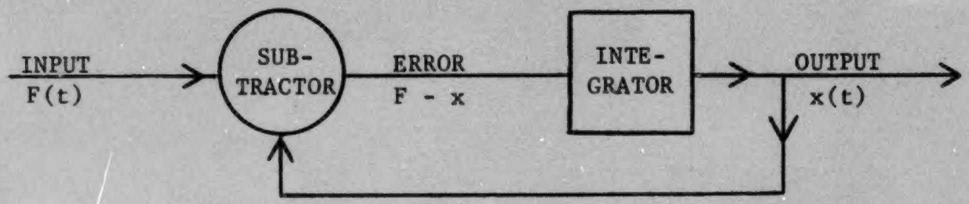
$$a(t)\, D_t x + x(t) = F(t)$$

as a model for various control problems. In this equation $F(t)$ is the input, $x(t)$ is the output, and $a(t)\, D_t x$ is an obstacle (causing a transient error) to achieving our desire that $x(t)$ closely approximate $F(t)$.

Suppose we assume a = positive constant. We may rewrite the differential equation as an integral equation

$$x = (1/a) \int (F - x)\, dt,$$

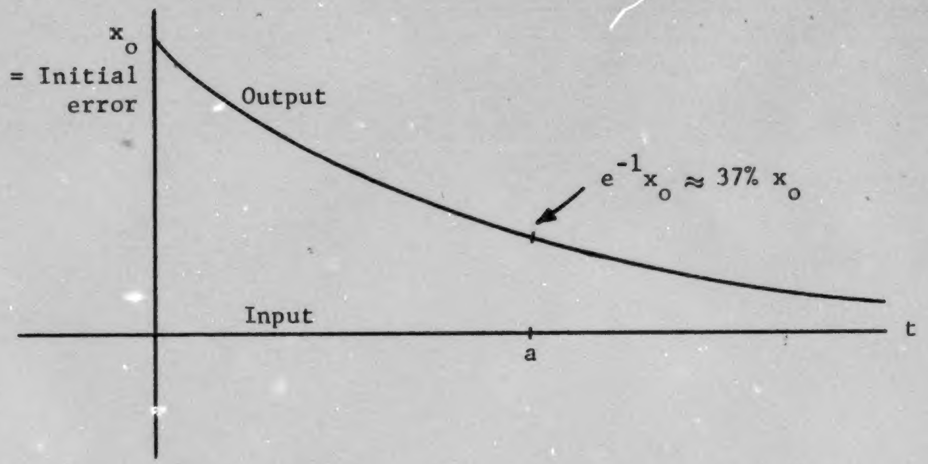
which has the following "block diagram" or "black box" representation:



For a = positive constant we will consider four cases: (i) $F(t) \equiv 0$, (ii) $F(t) \equiv$ a nonzero constant F_o, (iii) piecewise continuous inputs (primarily step functions), (iv) sinusoidal inputs.

Case 1. $F(t) \equiv 0$. Then the general solution is

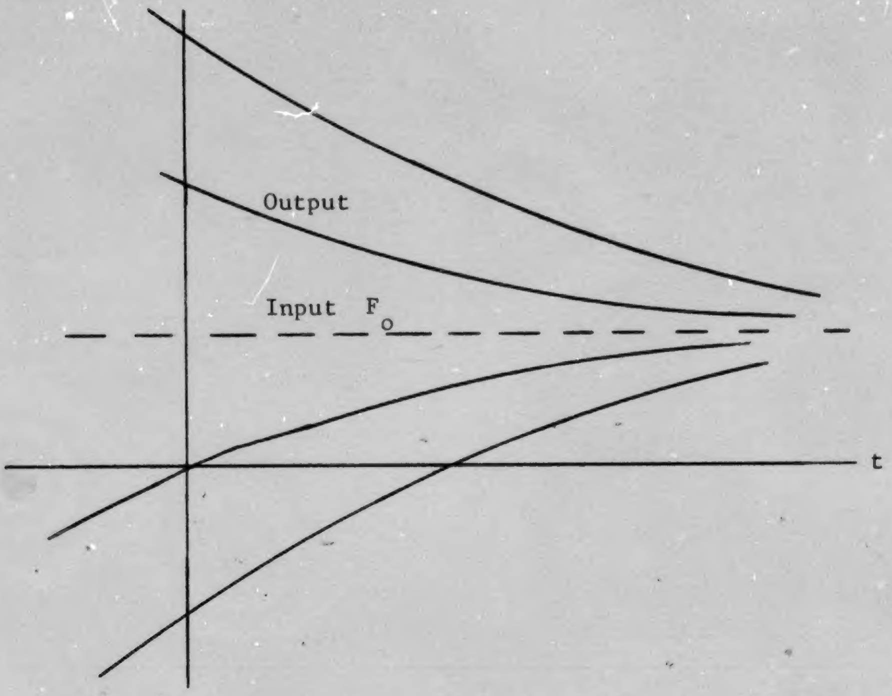
$$x = x_o e^{-t/a}.$$



The value a, called the time constant, is the time for the error to drop to 37% of its initial value. All of the solutions of this equation are transients, i.e., they go to zero as $t \to +\infty$. Hence the system is stable. For a = negative constant, we have an unstable system.

Case 2. $F(t) \equiv$ constant F_o. Then

$$x = F_o + ce^{-t/a}.$$

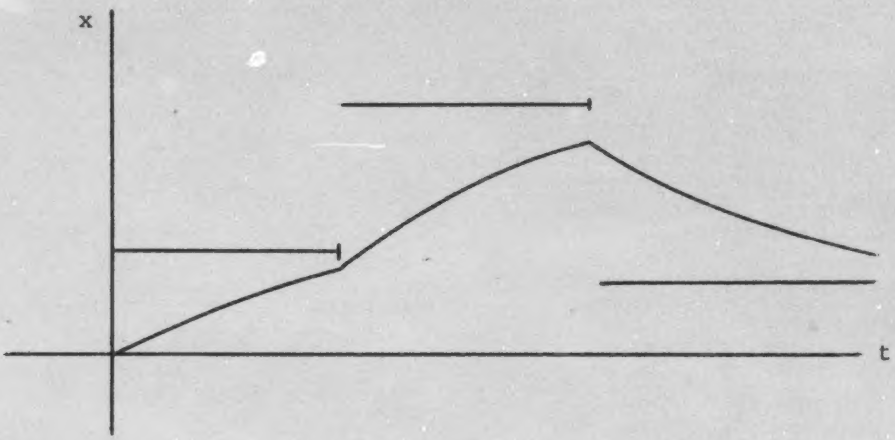


The effect of the constant F_o is to displace the equilibrium solution $x = 0$ to $x = F_o$.

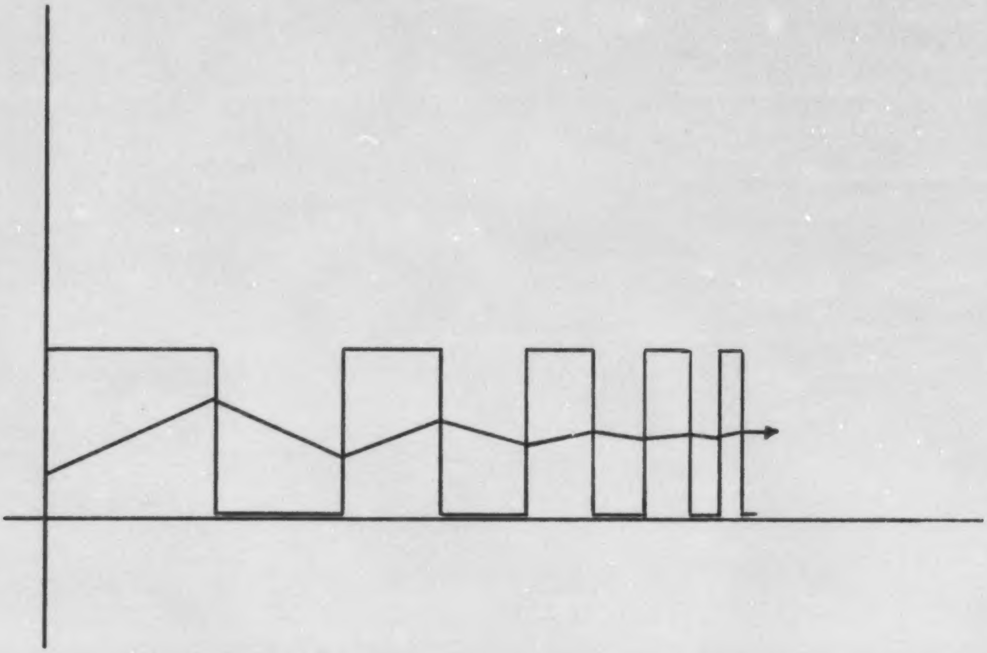
Case 3. Piecewise continuous input. The concept of piecewise continuous functions should be introduced early, for these functions have many applications and they demonstrate to a student that a function need not be described by a single analytic expression. If $F(t)$ has only finite jumps, then $G(t) = \int F(t)\,dt$ is well defined and $G'(t) = F(t)$ except at the jumps. The solution of $a\,Dx + x = F(t)$ by the usual formula gives

$$x = (1/a)\, e^{-t/a} \left(\int F(t)\, e^{t/a}\, dt + C \right),$$

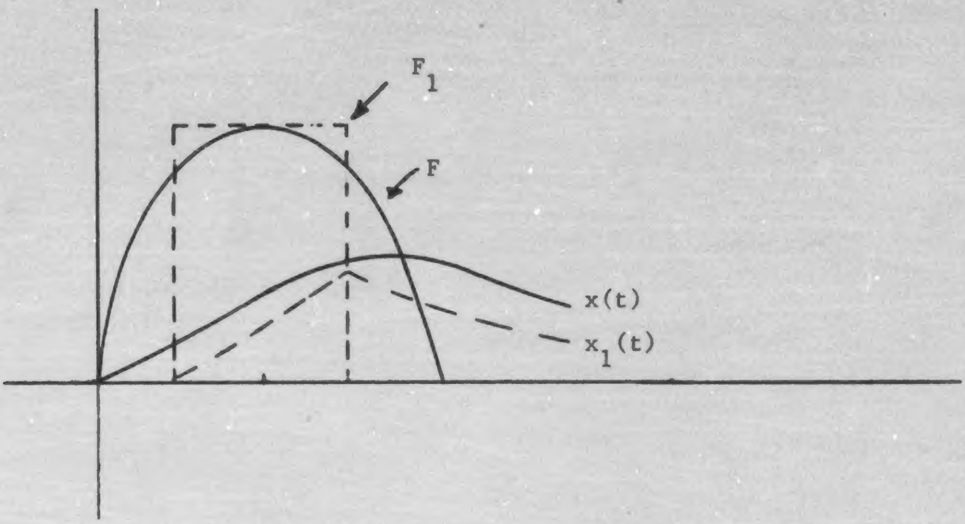
which satisfies the differential equation wherever F is continuous. If initial conditions are given, then this equation defines a unique continuous function which satisfies the differential equation wherever F is continuous. From the continuity of the solution it follows that the initial value problem for the case F = step function may be solved by treating it as a sequence of initial value problems for the case F ≡ constant. Pictorially:



If the input oscillates between two values (e.g., flicking a light switch) with increasing frequency, the output converges to a constant "average" value.



Approximating a continuous input by a step function can give a very good approximation to the output. For example, approximating a sinusoidal impulse by a suitable square impulse gives a very good approximation of the output.



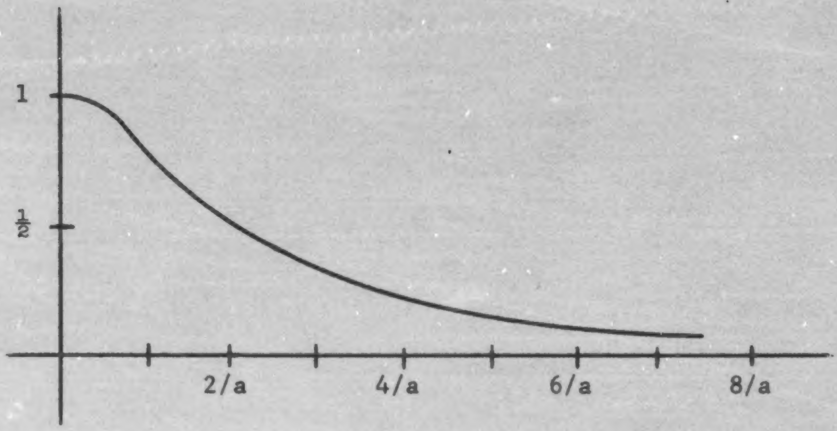
Also, we observe from graphs that if the input is a periodic step function, then the output is periodic (except for transient terms) [proof outlined below].

Case 4. Sinusoidal input. $a D_t x + x = A_i \sin \omega t$, A_i = input amplitude. Then the output is

$$x(t) = A_o \sin (\omega t - \alpha) + \text{transients},$$

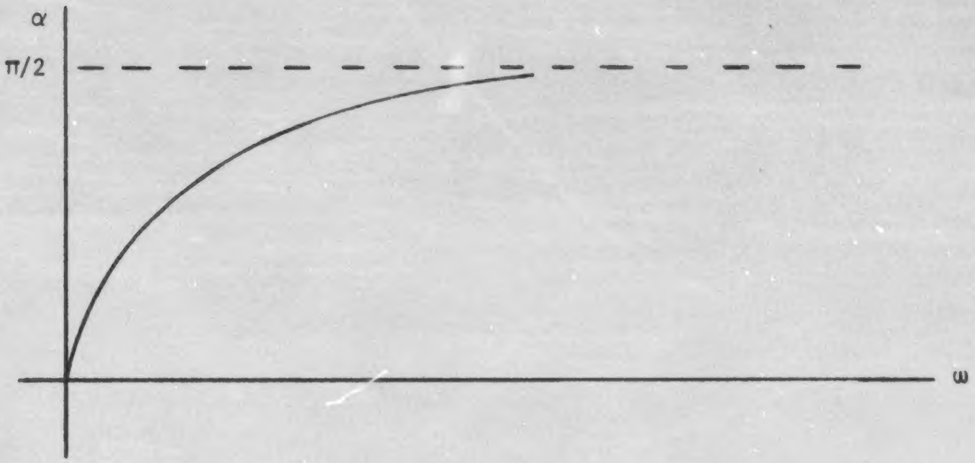
where $A_o = A_i / \sqrt{1 + a^2 \omega^2}$ and $\tan \alpha = a\omega$, $0 < \alpha < \pi/2$. We are assuming $a > 0$. If $a < 0$, we have an unstable system. Some examples of instability in applications are population growth and inflation.

Graphing the amplification factor A_o/A_i as a function of ω, we obtain:

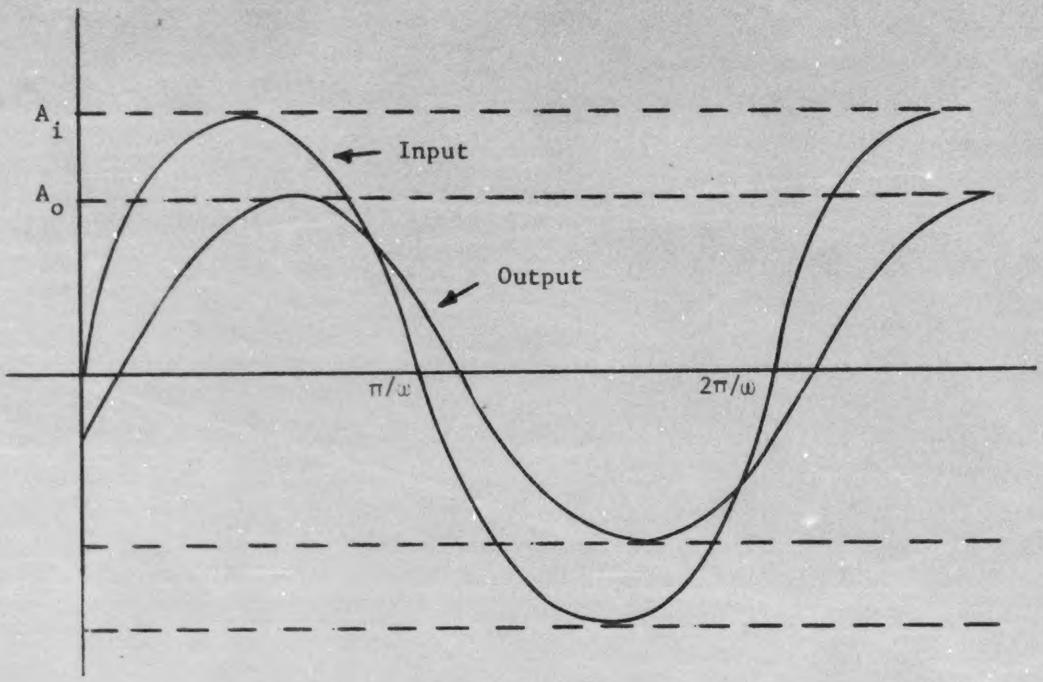


We see that the output amplitude A_o is less than the input amplitude A_i and that the amplification factor A_o/A_i decreases as ω increases.

Graphing the phase lag α as a function of ω, we obtain:



The phase lag and the amplitude suppression may be illustrated by showing the input and output as functions of time on the same graph.



One can use Fourier series to describe the behavior of the system for a general periodic F(t). It is appropriate to mention Fourier series in an elementary differential equations course. Although the students may not have had a formal treatment of Fourier series in a mathematics course, they have probably seen these series used in an engineering course.

For a periodic function F,

$$F(t) = \tfrac{1}{2}a_o + \sum_{n=1}^{\infty} (a_n \cos n\omega t + b_n \sin n\omega t),$$

$$\text{where } \omega = 2\pi/T.$$

For an input $A_i \sin(\omega t + \beta)$, the output is $A_o \sin(\omega t + \beta + \alpha)$, where A_o and α are given by the same formulas stated above for $A_i \sin \omega t$. Then rewriting

$$F(t) = \tfrac{1}{2}a_o + \sum_{n=1}^{\infty} A_n \sin(n\omega t + \beta_n),$$

we may obtain the output x(t) by "superposition":

$$x(t) = \tfrac{1}{2}a_o + \sum_{n=1}^{\infty} \frac{A_n}{\sqrt{1 + a^2 n^2 \omega^2}} \sin(n\omega t + \beta_n + \alpha_n),$$

where $\tan \alpha_n = an\omega$. Observe that the average of $F(t)$, $\frac{1}{2}a_o$, is not affected. Also, the amplitude factor decreases with n. Hence,

$$x(t) \sim \tfrac{1}{2}a_o + \frac{A_1}{\sqrt{1 + a^2 \omega^2}} \sin (\omega t + \beta_1 + \alpha_1).$$

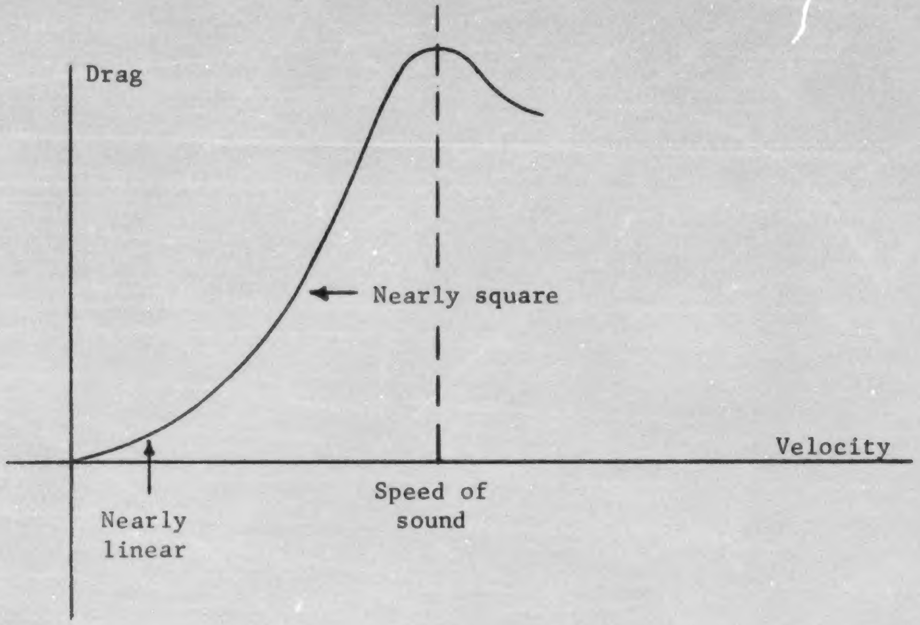
This explains why the output from a square wave input is a good approximation to the output from a sinusoidal input.

DIFFERENTIAL EQUATIONS

Lecture 3: Air Resistance, Higher-Order Systems

Air resistance is treated in books on ballistics. Also see Scarborough [28] and von Kármán and Biot [31].

Theoretical arguments for the kv^2 law go back as far as Newton, but various writers admit that "in fact" k depends on v. The general behavior is:



Reference: Kaplan [17], Sections 5-1 to 5-5, 5-9, 5-10, 5-12.

Higher-order systems:

$$a_o(t)\, D^n x + \dots + a_n(t)\, x = F(t), \quad a_o(t) \not\equiv 0.$$

If $a_{k+1}(t) \equiv \dots \equiv a_n(t) \equiv 0$ and $a_k(t) \not\equiv 0$, we may set $y = D^k x$ and obtain the differential equation

$$a_o(t)\, D^{n-k} y + \dots + a_k\, y = F(t).$$

We then obtain x from y by integration. This introduces a kind of instability, since x will be of the form

$$x(t) = g(t) + \text{polynomial in } t,$$

and the polynomial diverges to $\pm\infty$ as $t \to +\infty$.

If $a_n(t) \neq 0$ for all t in some neighborhood of t_o, then we may divide through by $a_n(t)$, obtaining a differential equation of the form

$$a_o(t)\, D^n x + \dots + a_{n-1}(t)\, Dx + x = F(t). \tag{1}$$

Now we see that $x(t) \sim F(t)$ if the coefficients $a_o(t), \dots, a_{n-1}(t)$ are small. Hence we may think of these coefficients as describing an obstacle which prevents the output $x(t)$ from being an exact copy of the input.

The solutions to differential equation (1) are of the form

$$x(t) = x^*(t) + c_1 x_1(t) + \dots + c_n x_n(t),$$

where $x_1(t), \dots, x_n(t)$ are solutions of the corresponding homogeneous equation.

The equation is stable if each $x_j(t)$ goes to zero as $t \to +\infty$. Then

$$x(t) = x^*(t) + \text{transients}.$$

If the coefficients are all constants, we will have stability if and only if all the roots of the characteristic equation $a_o r^n + \dots + a_n = 0$ have negative real parts.

If r is a simple root, then $\text{Re}(r) = 0$ if and only if the corresponding solutions are of the form $c_1 \cos \omega t + c_2 \sin \omega t$. Since these solutions are bounded, t[illegible] have neutral stability.

Liapounoff stability. For suitably small (bounded by $\delta > 0$) initial conditions, x and its derivatives $Dx, \dots, D^{n-1}x$ are bounded (by $\varepsilon > 0$) in the interval $0 \leq t < \infty$.

In an elementary course one should consider at least the second-order case. Then the roots of the characteristic equation

$$ar^2 + br + c = 0, \quad a > 0,$$

all have negative real parts if and only if $b > 0$ and $c > 0$.

The third-order case

$$ar^3 + br^2 + cr + d = 0, \quad a > 0,$$

is stable if and only if $b, c, d > 0$ and $bc > ad$.

The n^{th}-order case is solved by the Hurwitz-Routh criterion (see Gantmacher [12]). A finer result is given by the Liénard-Chippart theorem.

Second-order equations with constant coefficients. Consider

$$a_oD^2x + a_1Dx + x = F(t),$$

where a_o, a_1, a_2 are positive constants. The corresponding homogeneous equation may be written as

$$D^2x + 2h\,Dx + \lambda^2 x = 0,$$

by setting $h = a_1/2a_o$ and $\lambda = 1/\sqrt{a_o}$. Then the roots of the characteristic equation are $-h + \sqrt{h^2 - \lambda^2}$ and $-h - \sqrt{h^2 - \lambda^2}$. We have three cases:

1. $0 < h < \lambda$ "underdamped"

   $x = Ae^{-ht}(\sin \beta t + \alpha)$

2. $h = \lambda$ "critical damping"

   $x = e^{-ht}(c_1 + c_2 t)$

3. $h > \lambda$ "overdamped"

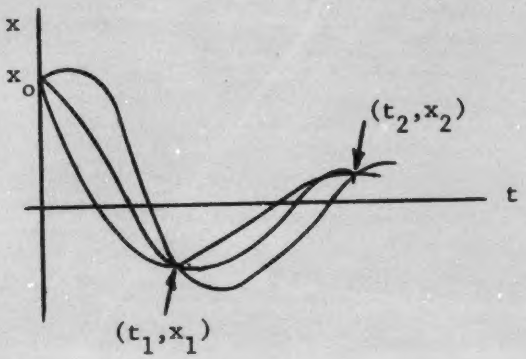
   $x = c_1e^{-at} + c_2e^{-bt}, \quad a > b > 0,$

   $a = h + \sqrt{h^2 - \lambda^2}, \quad b = h - \sqrt{h^2 - \lambda^2}.$

Case 2 is less important in practice since it is much less likely to occur. Case 3 resembles the first-order case.

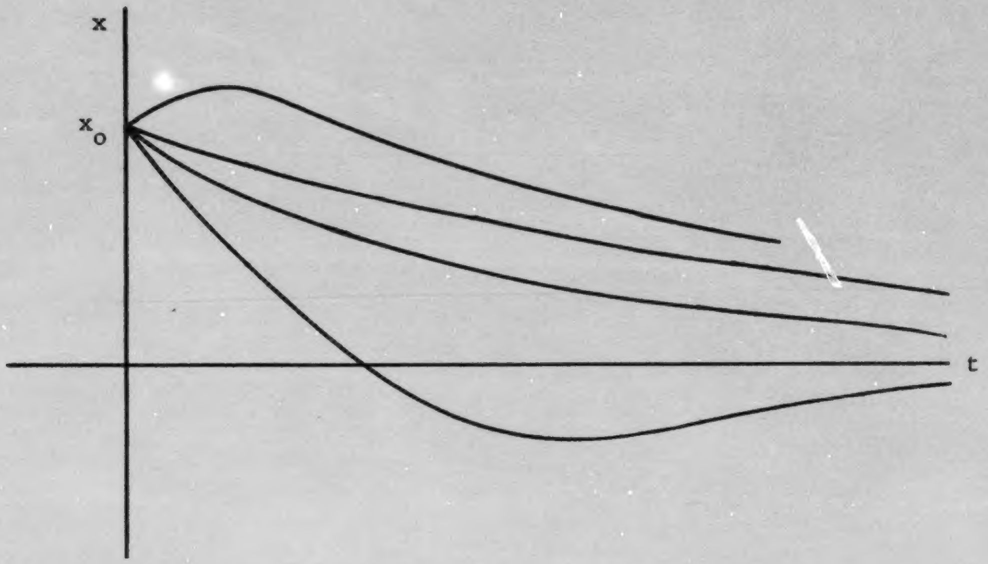
Students should become thoroughly familiar with the graphs.

Keeping x_o fixed but allowing v_o to vary, one obtains the following for Case 1:



As v_o increases, the scale increases and the graph shifts to the right; but there is a sequence of points in the plane through which all graphs pass. That is, there is a time $t_1 > 0$ and a corresponding x_1 such that at the time t_1 the system will be at $x = x_1$ regardless of the initial velocity v_o. Similarly, there are common points (t_2,x_2), (t_3,x_3), etc., with $t_1 < t_2 < t_3 < \ldots$.

For Case 3, fixing x_o and varying v_o one has:



For cases 1 and 2 the time constant τ (which is a measure of the rate at which the transient approaches zero) is taken to be

$$\tau = 1/h = 2\ a_o/a_1.$$

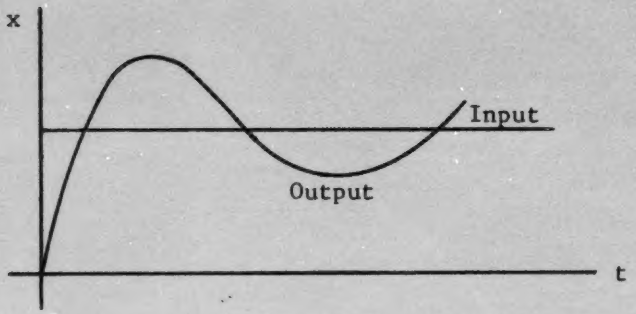
For case 3 the time constant is taken from the larger exponential factor e^{-bt}. Hence

$$\tau = \frac{1}{h - \sqrt{h^2 - \lambda^2}} = \frac{2a_o}{a_1 - \sqrt{a_1^2 - 4a_o}}.$$

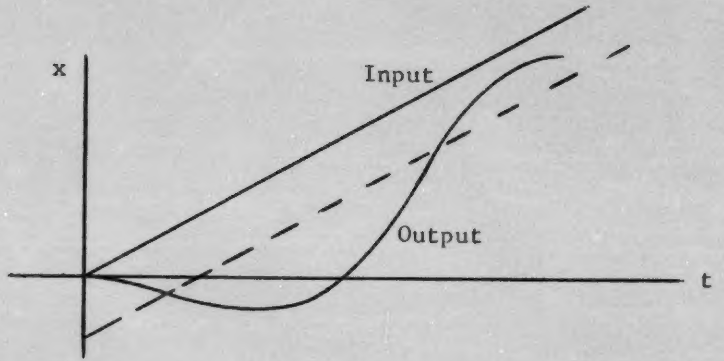
Note that if a_o is very small (relative to a_1), we get an approximation to the time constant for the first-order case, i.e.,

$$\tau = \frac{2a_o}{a_1 - \sqrt{a_1^2 - 4a_o}} \quad \to \ a_1 \ \text{ as } \ a_o \to 0.$$

Considering the outputs for various inputs for case 1, we see:



Response to constant input.



Response to ramp (linear) input.

As in the first-order case, one can generalize from constant inputs to step-function inputs by requiring that x(t) and x'(t) be continuous and decomposing the problem into a sequence of initial value problems. Observe that one can build up the amplitude by a proper timing of the inputs as, for example, one does in pushing a child on a swing; this is an example of "resonance."

Second-order equations with variable coefficients. When is the equation

$$a_o(t)\, D^2x + a_1(t)\, Dx + a_2(t)\, x = 0$$

stable? A general treatment is very difficult. From the constant coefficient case, one might expect that $a_o(t)$, $a_1(t)$, $a_2(t) > 0$ for all t would imply stability, but this is not the case. For example, the general solution of the "Euler" equation

$$6t^2 D^2x + t\,Dx + x = 0$$

is $x = c_1 t^{1/3} + c_2 t^{1/2}$. Hence the equation

$$6t^2 D^2x + t\,Dx + x = F(t)$$

is unstable.

Another approach is to consider the roots of the characteristic equation as functions of time. That this is not always reliable may be shown by Mathieu's equation

$$D^2x + (a - b \cos 2\omega t)x = 0,$$

which is exponentially unstable even with $|b| < |a|$.

There are results which depend upon the boundedness of the coefficients, e.g., the

Perron Stability Theorem. Given

$$\frac{d\vec{x}}{dt} + A(t)\,\vec{x} = \vec{F}(t). \qquad (*)$$

If all entries in the matrix $A(t)$ are continuous and bounded for $0 \leq t < \infty$, then the solutions of

$$\frac{d\vec{x}}{dt} = A(t)\,\vec{x}$$

are "exponentially stable," i.e.,

$$\|\vec{x}\| \leq k\,\|\vec{x}_o\|\,e^{-\lambda t}$$

for some $k, \lambda > 0$ if and only if every solution of (*) is bounded for every bounded input vector function $\vec{F}(t)$.
[See Kaplan [16].]

Examples of physical applications

(a) Mass suspended by a spring with or without an added dashpot

(b) Simple pendulum

(c) Liquid in a U-shaped glass tube [Phillips [25], p. 113, problem 11.]

(d) Rod resting on a cylinder [Scarborough [28], p. 231, or Phillips [25], _loc_. _cit_., problem 12.]

(e) Electric circuits [will be treated below]

DIFFERENTIAL EQUATIONS

Lecture 4: Existence Theory

Existence theory--linear vs. nonlinear equations. For a linear differential equation

$$a_o(t)\, D^n x + \dots + a_n(t)\, x = F(t),$$

if $a_o(t) \neq 0$ and if $F(t)$ and all of the $a_j(t)$ are continuous for all t in some interval $a < t < b$, then for given initial conditions there exists a unique solution over the whole interval. For nonlinear cases, on the other hand, the existence and uniqueness may only be local. For a first-order equation

$$\frac{dx}{dt} = f(x,t),$$

if $f(x,t)$ and $f_x(x,t)$ are continuous for all x and for all t in some interval $a < t < b$, [Note. For a particular $f(x,t)$ it is usually more convenient to verify the condition that f_x is continuous than it is to verify a Lipschitz condition.] then one can extend a solution through a given point (t_o, x_o) to the boundary of the region $a < t < b$, $-\infty < x < +\infty$; however, it may diverge to the "top" or "bottom" of the region before it reaches the sides. If this is the case, the differential equation is said to have finite escape time.

For example, if $f(x,t) = 1 + x^2$, then f and f_x are continuous for all x and for all t. Hence one may take $-\infty < t < +\infty$. One can easily show that $x = \tan t$ is the unique solution of $\frac{dx}{dt} = 1 + x^2$ passing through (0,0), but $\tan t \to +\infty$ as $t \to \pi/2$.

Stability of equilibrium for nonlinear systems. Consider the autonomous case (i.e., each f_i is independent of t)

$$\frac{dx_i}{dt} = f_i(x_1, \dots, x_n), \quad i = 1, \dots, n.$$

Suppose there exists an equilibrium point (also called a critical or stagnation point) $(x_1^0, \dots, x_n^0)$ such that $f_i(x_1^0, \dots, x_n^0) = 0$ for $i = 1, \dots, n$. Then $x_i(t) \equiv x_i^0$ is the equilibrium solution. We translate the origin to the equilibrium point and then expand f_i in power series, obtaining

$$\frac{dx_i}{dt} = \sum a_{ij}\, x_j + \text{higher-order terms, where } a_{ij} = \partial f_i/\partial x_j.$$

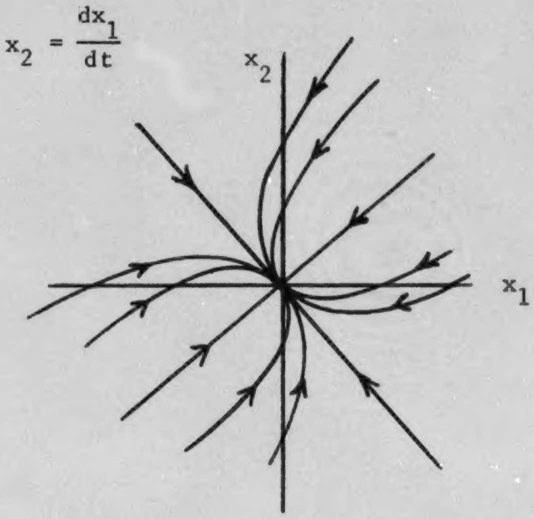
We then look at the linear approximation

$$\frac{dx_i}{dt} = \sum a_{ij}\, x_j.$$

Theorem. If $\frac{d\vec{x}}{dt} = A\,\vec{x}$ has $\vec{x} \equiv 0$ as an asymptotically stable solution, then so does $\frac{d\vec{x}}{dt} = \vec{f}(\vec{x})$.

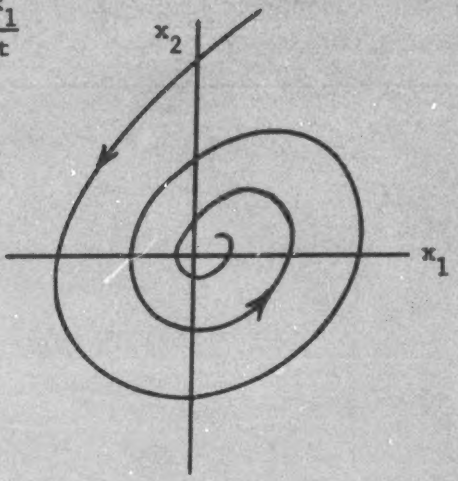
The possible behavior is illustrated by the following graphs in the phase plane:

Structurally stable (i.e., not changed by linearization):



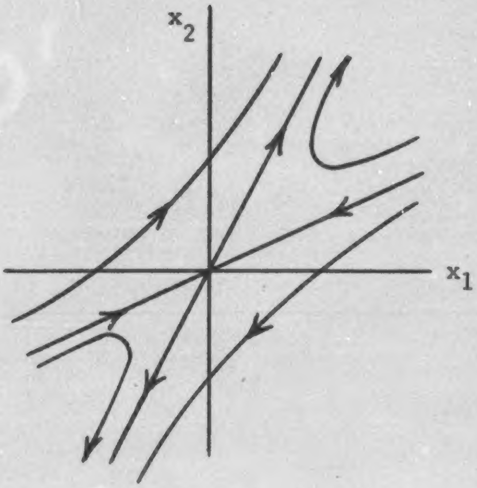
Node (stable)

$$x_2 = \frac{dx_1}{dt}$$

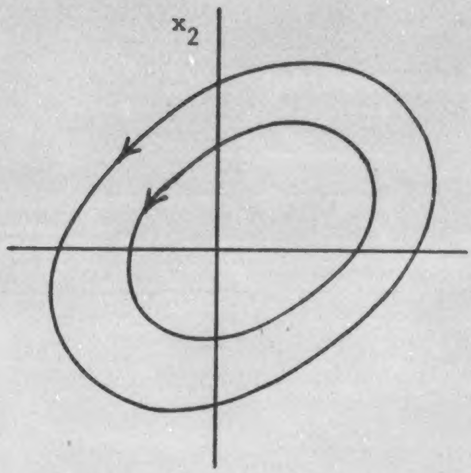


Focus (stable)

Structurally unstable:



Saddle (unstable)

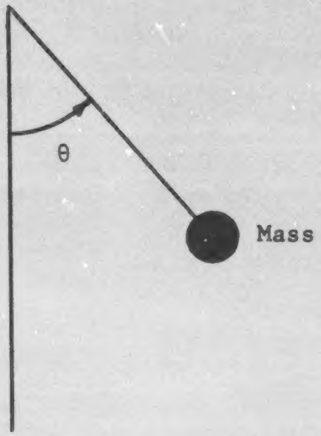


Center (neutrally stable)
(slight change can destroy
geometric pattern)

Simple harmonic motion is a familiar example of a center pattern.

In many applications damping is ignored, linearization thus leading to the center pattern; hence, there is a question of whether or not the correct pattern is obtained.

Pendulum:



Ignoring damping, the equation of motion is

$$\frac{d^2\theta}{dt^2} = -k^2 \sin \theta .$$

Observe that $\theta \equiv 0$ is a solution. Substituting

$$x_1 = \theta \text{ and } x_2 = \frac{d\theta}{dt}$$

and linearizing, we have

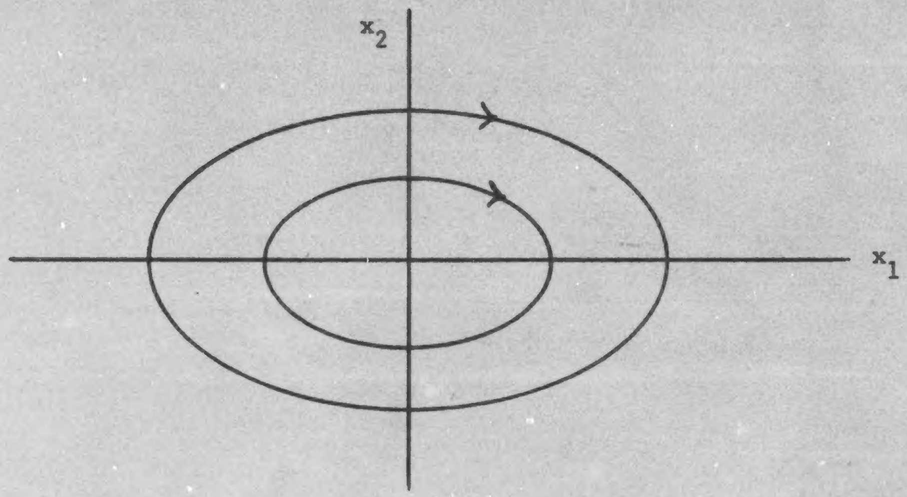
$$\frac{dx_1}{dt} = x_2 \text{ and } \frac{dx_2}{dt} = -k^2 x_1 .$$

Eliminating t gives

$$\frac{dx_2}{dx_1} = -k^2 \frac{x_1}{x_2},$$

whose integral curves $x_2^2 + k^2 x_1^2 = C$ are ellipses with center at the origin.

Hence one obtains a center.



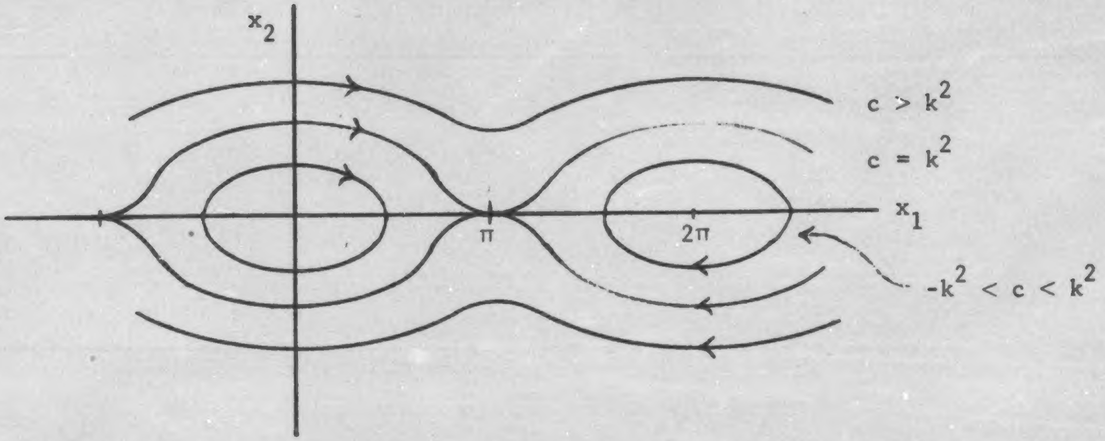
In this example the original equation also has a center at the origin in the x_1,x_2-diagram. Integrating $d^2\theta/dt^2 = -k^2 \sin \theta$ gives

$$\frac{1}{2} \left(\frac{d\theta}{dt}\right)^2 = k^2 \cos \theta + C.$$

After substituting $x_1 = \theta$, $x_2 = d\theta/dt$, one obtains

$$(x_2)^2 = 2k^2 \cos x_1 + 2C.$$

Graphing:



There are centers at $(0,0)$, $(2\pi,0)$, etc., and saddles at $(\pi,0)$, $(3\pi,0)$, etc. One can show that for $C = -k^2$ uniqueness implies that it takes infinite time to reach a saddlepoint.

It can be shown that for nonlinear equations of the form $d^2x/dt^2 = f(x)$ one does get a correct pattern at the centers by linearization. [This is discussed in more detail in the lectures by Professor Robert Ritt].

Systems of second-order linear differential equations. One method of solving

$$D^2 x_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, n$$

is to make an *exponential substitution*

$$\vec{x} = e^{\lambda t} \vec{\mu},$$

where $e^{\lambda t}$ is a scalar function and $\vec{\mu}$ is a constant vector, obtaining

$$\det(A - \lambda^2 I) = \begin{vmatrix} a_{11} - \lambda^2 & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda^2 & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda^2 \end{vmatrix} = 0;$$

the right-hand side is a polynomial in λ^2 of degree n. Suppose we have distinct nonzero real roots $\lambda^2 = \gamma_1, \ldots, \gamma_n$. Then we have 2n distinct values for λ, namely

$$\lambda_1 = \sqrt{\gamma_1},\ \lambda_2 = -\sqrt{\gamma_1},\ \lambda_3 = \sqrt{\gamma_2},\ \ldots,\ \lambda_{2n} = -\sqrt{\gamma_n},$$

and corresponding eigenvectors $\vec{\mu}_1, \ldots, \vec{\mu}_{2n}$.

Or if all the γ's are negative, then the values for λ are of the form $\pm\omega_1 i, \ldots, \pm\omega_n i$, and one obtains the general solution

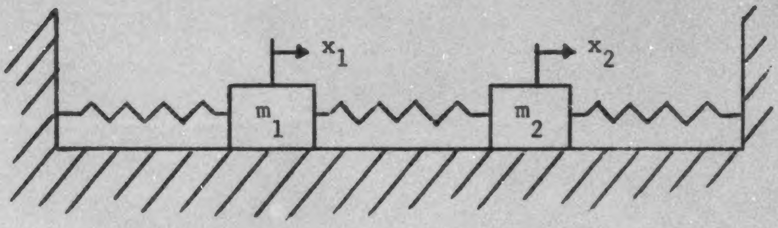
$$\vec{x} = C_1 \operatorname{Re}(e^{i\omega_1 t})\vec{\mu}_1 + \ldots + C_n \operatorname{Re}(e^{i\omega_n t})\vec{\mu}_n$$

$$+ C_{n+1} \operatorname{Im}(e^{i\omega_1 t})\vec{\mu}_1 + \ldots + C_{2n} \operatorname{Im}(e^{i\omega_n t})\vec{\mu}_n,$$

which is neutrally stable.

In a physical problem if one gets repeated roots, then one can still pick linearly independent eigenvectors. Hence the solution is similar to that of the distinct root case.

Example. Consider two masses m_1 and m_2 held by 3 identical springs:



Introduce coordinates x_1 and x_2 relative to the equilibrium positions. Then (assuming no friction)

$$m_1 D^2 x_1 = -2k^2 x_1 + k^2 x_2$$

$$m_2 D^2 x_2 = k^2 x_1 - 2k^2 x_2.$$

By exponential substitution one obtains a solution of the form:

$$\lambda = \pm\omega_1 i, \pm\omega_2 i, \quad 0 < \omega_1 < \omega_2$$

$$x_1 = k^2 A_1 \sin (\omega_1 t + \phi_1) + k^2 A_2 \sin (\omega_2 t + \phi_2)$$

$$x_2 = (2k^2 - m_1\omega_1^2) A_1 \sin (\omega_1 t + \phi_1) + (2k^2 - m_1\omega_2^2) A_2 \sin (\omega_2 t + \phi_2).$$

If $A_2 = 0$, then the two masses are oscillating with the same frequency ω_1, with the same phase ϕ_1 (perfect sychronization), and with amplitudes in a fixed ratio. Similarly if $A_1 = 0$, the two masses are oscillating in sychronization at the common frequency ω_2. These two types of sinusoidal motion are called the normal modes of the physical system. The general motion is a linear combination of these two modes.

This type of analysis has further application in the theory of molecular spectra.

DIFFERENTIAL EQUATIONS

Lecture 5: Electrical Circuits and Networks

Lagrange equations. [Reference: Kaplan [17], Section 6-13.] The differential equations for unforced small oscillations of a conservative system about an equilibrium can be written in the Lagrangian form:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial q_i'}\right) - \frac{\partial L}{\partial q_i} = 0, \; i = 1, \ldots, n,$$

where $q_1, \ldots, q_n$ are coordinates which determine the system at time t, chosen so that $(q_1, \ldots, q_n) = (0, \ldots, 0)$ at equilibrium, $q_i' = dq_i/dt$, and $L = T - V$, where T is the total kinetic energy and V is the total potential energy of the system. The function L is called the Lagrangian of the system. [See books on mechanics, e.g., Goldstein [13].]

In the problem, considered above, of two weights restrained by three springs,

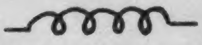
$$T = \tfrac{1}{2}\left[m_1(dx_1/dt)^2 + m_2(dx_2/dt)^2\right]$$

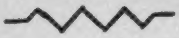
and

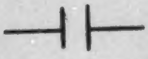
$$V = \sum b_{ij}x_i x_j,$$

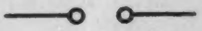
where $q_1 = x_1$, $q_2 = x_2$, $q_1' = dx_1/dt$, $q_2' = dx_2/dt$.

Electrical circuits and networks. [Reference: Kaplan [17], Section 6-14.] One is concerned with the behavior of electrical current subject to the parameters:

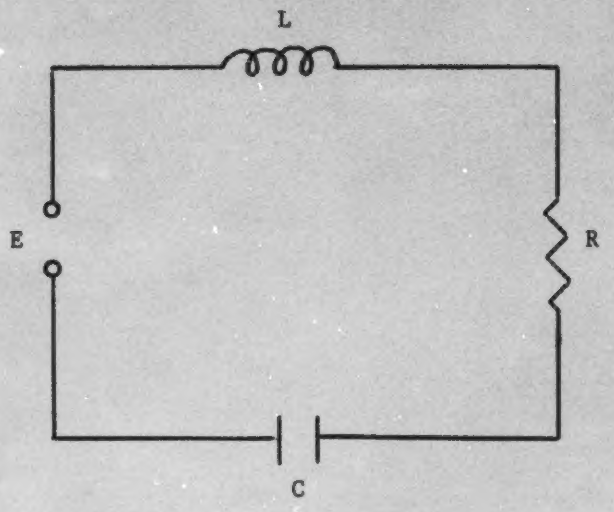
L = inductance (magnetic effect) 

R = resistance (heating) 

C = capacitance (condenser) 

E = applied voltage 

These may all appear in a simple circuit:



The current in various branches of a network are subject to the Kirchhoff Laws:

I. The net current flowing into a node is zero. Hence, at a given time the current I is the same at all points of a branch.

II. The total voltage around each mesh (circuit) is zero.

For (I) we require: If q is the charge on a condenser, then dq/dt = I.

For (II) we require: The voltage drop due to:

an inductance L is L(dI/dt),
a resistance I is RI,
a capacitance C is q/C,
an applied voltage E is -E.

Hence, the differential equation of a simple electric circuit is

$$L\frac{dI}{dt} + RI + \frac{q}{C} - E = 0,$$

which may be written as

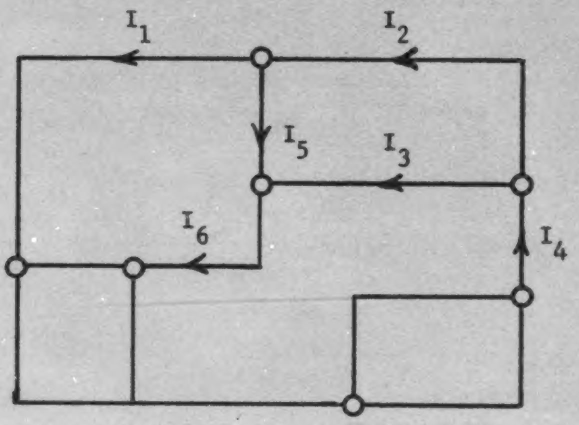
$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = E$$

or as

$$L\frac{dI}{dt} + RI + \frac{1}{C}\int I\,dt = E.$$

Normally, L, R, and C are assumed constant, but E may be a function of time.

In the following network, for example,



Kirchhoff's first law gives linear algebraic relationships among the currents at each node. If there are N nodes (vertices) and B branches (also called arcs or edges), then there are N linear algebraic equations restricting the B currents $I_1, I_2, \ldots, I_B$. In the example above we have:

$$-I_1 + I_2 - I_5 = 0$$

$$I_3 + I_5 - I_6 = 0$$

$$-I_2 - I_3 + I_4 = 0,$$

etc., where each equation describes the net current entering a particular node.

Of these N equations, at most N - 1 are independent. For if we add the left members of the first N - 1 equations we have:

$$0 = \Sigma \text{ (all currents entering nodes 1 through } N - 1)$$

$$= \Sigma \text{ (currents entering nodes 1 through } N - 1 \text{ from nodes 1 through } N - 1) + \Sigma \text{ (currents leaving node } N).$$

But

$$\Sigma \text{ (currents entering nodes 1 through } N - 1 \text{ from nodes 1 through } N - 1) = 0,$$

since each branch appears twice, once with a plus sign and once with a minus sign. Hence

$$0 = \Sigma \text{ (currents leaving node N)}$$

$$= -\Sigma \text{ (currents entering node N)}.$$

Hence the first N - 1 equations imply the N^{th} equation. Therefore: The set of possible values of the branch currents $(I_1, \ldots, I_B)$ is a subspace of a B-dimensional real vector space E^B of dimension $\geq B - (N - 1) = B - N + 1$.

Applying Kirchhoff's second law, one obtains a linear differential equation for each circuit, for example

$$L_2\, dI_2/dt + R_2 I_2 + (q_2/C_2)$$

$$- L_3\, dI_3/dt - R_3 I_3 - (q_3/C_3)$$

$$+ L_5\, dI_5/dt + R_5 I_5 + (q_5/C_5) = 0,$$

where L_2 is the inductance, R_2 is the resistance, and C_2 is the capacitance in branch 2 and q_2 is the charge on the condenser in branch 2, etc. For a planar network, it is sufficient to take only those meshes which enclose a single bounded region, obtaining M equations, where M is the number of bounded regions. Adding 1 for the region outside the network, Euler's formula is

$$N - B + (M + 1) = 2.$$

Hence,

$$(N - 1) + M = B$$

and

$$B - N + 1 = M.$$

This is consistent with our claim that there are at most N - 1 independent algebraic equations and M independent differential equations.

The computation of the solutions of such a system can be very tedious. It can be simplified by using the concept of a mesh current. We can associate with each mesh i a unique counterclockwise "current" J_i such that if I is the current in a branch which is a common edge of meshes p and q, then $I = \pm(J_p - J_q)$, where the choice of sign is determined by the direction of the current I. We take the current of the mesh bounding the unbounded region surrounding the network to be $J_{ext} = 0$.

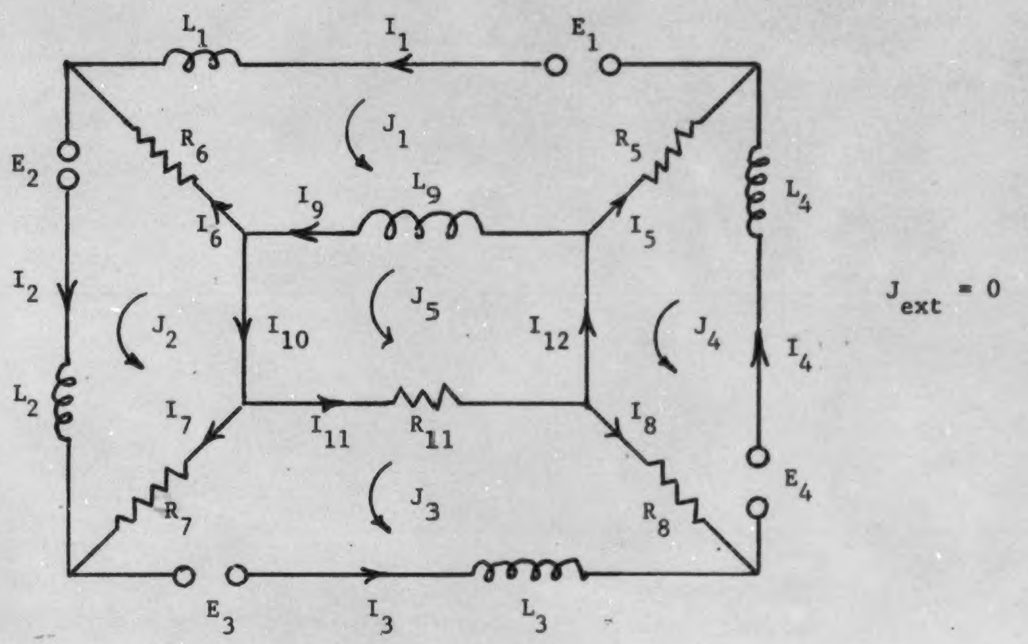
In general, we have mesh currents $J_1, \ldots, J_M$ and branch currents $I_1, I_2, \ldots, I_B$ such that

$$I_\alpha = \sum C_{\alpha\beta} J_\beta, \text{ where } C_{\alpha\beta} = -1, 0, \text{ or } +1.$$

These equations constitute a linear mapping Γ from an M-dimensional real vector space E^M into a B-dimensional real vector space E^B. What is the kernel of this mapping? Suppose all of the branch currents are zero. If $J_p = 0$ and the p^{th} and q^{th} mesh have a branch in common, then $0 = J_p - J_q$. Hence $J_q = 0$. It follows from the finiteness of our network that if one mesh current is zero (assuming all of the branch currents are zero), then they all are. But $J_{ext} = 0$. Therefore all the J's are zero. That is, the kernel of the mapping Γ is (0, 0, ..., 0). Hence Γ is a one-to-one mapping from E^M into E^B. Therefore:

The set of possible values of the branch currents $(I_1, \ldots, I_B)$ is a subspace of E^B of dimension $M = B - N + 1$. Hence, N - 1 of the N current equations which follow from Kirchhoff's first law are independent.

Example. All capacitances = 0.



Express the branch currents in terms of the mesh currents:

$$I_1 = J_1, \ldots, I_4 = J_4$$

$$I_5 = J_1 - J_4, \; I_6 = J_2 - J_1$$

$$I_7 = J_3 - J_2, \; I_8 = J_4 - J_3$$

$$I_9 = J_5 - J_1, \; I_{10} = J_5 - J_2$$

$$I_{11} = J_5 - J_3, \; I_{12} = J_5 - J_4$$

Solve for the mesh currents, obtaining:

$$J_1 = I_1, \ldots, J_4 = I_4$$

$$J_5 = I_1 + I_9 \; (= I_2 + I_{10} = I_3 + I_{11} = I_4 + I_{12})$$

Now we write the differential equation for each mesh:

(1) $L_1 DI_1 - R_6 I_6 - L_9 DI_9 + R_5 I_5 = E_1$

(2) $L_2 DI_2 \ldots$, etc.

Next express the branch currents $I_1, \ldots, I_{12}$ in terms of the mesh currents $J_1, \ldots, J_5$, obtaining 5 linear differential equations in 5 unknown dependent variables--the mesh currents:

(1) $L_1 DJ_1 - R_6(J_2 - J_1) - L_9 D(J_5 - J_1) + R_5(J_1 - J_4) = E_1$

(2) $L_2 DJ_2 \ldots$, etc.

[Observe that by introducing mesh currents, we are able to replace a system of 7 (= N - 1) algebraic equations and 5 (= M) differential equations in 12 (= B) unknowns by a system of 5 differential equations in 5 unknowns.]

For the homogeneous case--$E_i = 0$ for all i--one can solve by an exponential substitution $J_i = e^{\lambda t} u_i$ or by reducing the equations to the form

$$dJ_k/dt = \sum a_{k\ell} J_\ell .$$

Engineers are primarily interested in the eigenvalues--the λ's.

If the network involves capacitances, then we introduce mesh "charges"

$$Q_\alpha = \int J_\alpha \, dt,$$

244

obtaining a system of differential equations of the form

$$\sum_{\beta=1}^{M} (L_{\alpha\beta}D^2 + R_{\alpha\beta}D + \gamma_{\alpha\beta})Q_\beta = E_\alpha, \ \alpha = 1, \ldots, M.$$

One can show that the matrices $[L_{\alpha\beta}]$, $[R_{\alpha\beta}]$, $[\gamma_{\alpha\beta}]$ are all symmetric--the Reciprocity Law.

If the resistances and applied voltages are all zero, then these equations can be written in the Lagrangian form:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial Q_\alpha'}\right) - \frac{\partial L}{\partial Q_\alpha} = 0,$$

where

$$Q_\alpha' = dQ_\alpha/dt,$$

$$L = T - V,$$

$$T = \underline{\text{total electromagnetic energy}}$$

$$= \tfrac{1}{2}\sum_{\alpha\beta} L_{\alpha\beta}Q_\alpha' Q_\beta',$$

$$V = \underline{\text{total electrostatic energy}}$$

$$= \tfrac{1}{2}\sum_{\alpha\beta} \gamma_{\alpha\beta}Q_\alpha Q_\beta.$$

If $R_{\alpha\beta}$ and E_α are not all zero, then we must also consider the dissipation function

$$F = \tfrac{1}{2}\sum R_{\alpha\beta}Q_\alpha' Q_\beta'.$$

It can be shown that T, V, $F \geq 0$ (and usually > 0) for all values of the variables, which implies that there is some form of stability--the Stability of Passive Networks Theorem.

DIFFERENTIAL EQUATIONS

Lecture 6: Differential Equations in Economics

Reference: Allen [2].

Mathematics in economics is in a state of flux. The teacher of differential equations needs to understand how well it is grounded and how widely it is accepted. The social sciences are relatively new in comparison to the physical sciences, and the use of mathematics in the social sciences is very recent. Some mathematics is now being used in archeology, anthropology, and history as well as in psychology, etc. To understand the present state of mathematics in the social sciences one must be in contact with people in those fields. The various fields tend to split into mathematical and nonmathematical schools.

J. M. Keynes used mathematics very heavily in economics and he had considerable influence on the young economists of his day. The economists tend to use engineering terminology--input, output, control, stability, lag, etc.

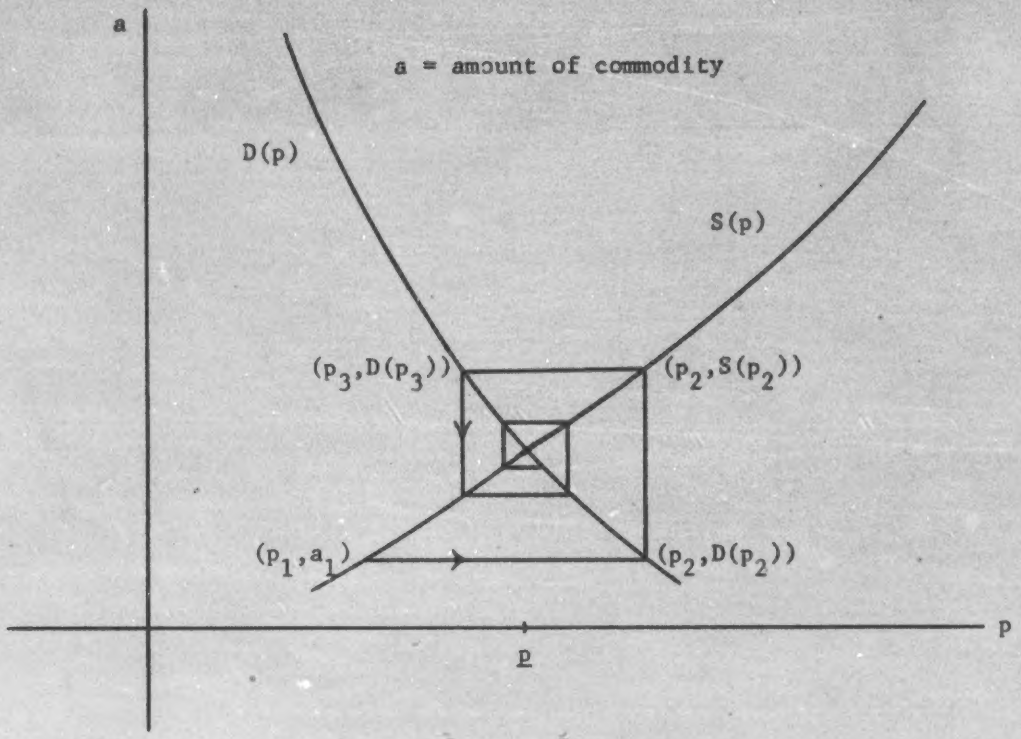
Theory of demand and supply. Consider one commodity which is for sale at a price p. Denote

$D(p)$ = *demand function*
= amount (in appropriate units) the buyers are willing to purchase at price p.
$S(p)$ = *supply function*
= amount offered for sale when the price is p.

One expects D to be decreasing and S to be increasing. The *equilibrium price* is the price $\underline{p}$ such that

$$D(\underline{p}) = S(\underline{p}).$$

Suppose the sellers offer a certain amount a_1 of the commodity at a price p_1, $p_1 < \underline{p}$, $a_1 < D(p_1)$. The demand raises the price to p_2 (see graph). The sellers then make an amount $S(p_2)$ available, which will depress the price to p_3, etc.



Depending on the shape of the demand and supply curves near the equilibrium, the path may spiral inward as shown (stable), oscillate (neutral), or spiral outward (unstable). This is referred to as the "cobweb model."

If the demand and supply curves are linear, then stability depends on the comparative steepness of the two curves:

$$|dD/dp| > |dS/dp| \quad \text{(stable)}$$

$$|dD/dp| = |dS/dp| \quad \text{(neutral)}$$

$$|dD/dp| < |dS/dp| \quad \text{(unstable)}$$

For a more sensitive model, assume that

$$D(p) = \alpha_1 + \alpha_2 p + \alpha_3 (dp/dt),$$

$$S(p) = \beta_1 + \beta_2 p,$$

and that the price adjusts to make D = S. Setting

$$\alpha_3 (dp/dt) + (\alpha_2 - \beta_2) p = \beta_1 - \alpha_1,$$

we obtain

$$p = \bar{p} + ce^{-t/\tau},$$

where $\bar{p} = (\beta_1 - \alpha_1)/(\alpha_2 - \beta_2)$, which must be > 0, and $(1/\tau) = (\alpha_2 - \beta_2)/\alpha_3$. If $(1/\tau) > 0$, then price stabilizes.

Domar model for "the burden of the national debt." Let

$D(t)$ = total national debt at time t,

$y(t)$ = total national annual income (dollars per year).

Assume

$$D(t + \Delta t) - D(t) \sim ky(t)\,\Delta t.$$

Hence

$$dD/dt = ky.$$

If $y = a + bt$, then $D \sim bt^2/2$ and $D/y \to \infty$, which is bad. But if $y = ce^{bt}$, then $D = \alpha + (kc/b)e^{kt}$ and $D/y \to k/b$. Hence, according to the model, the "burden" of the debt, D/y, can be kept under control if annual income increases exponentially ("geometrically") but not if it increases linearly ("arithmetically").

Problem. Apparently, economists have only considered two possibilities--y(t) is linear or y(t) is exponential. For a more thorough analysis one might let $u = D/y$. Then

$$\frac{du}{dt} = \frac{y\,(dD/dt) - D(dy/dt)}{y^2}$$

$$= \frac{yky - yu(dy/dt)}{y^2}$$

$$= k - \frac{u}{y}(dy/dt).$$

One should then try to characterize y(t) so that u is bounded.

Harrod-Domar investment model. Let

$y(t)$ = rate of income at time t,

$S(t)$ = rate of saving at time t.

Assume $S(t) = sy(t)$, where s = constant > 0. Let

$I(t)$ = desired rate of investment.

Assume $I(t) = g(dy/dt)$, where $g = \text{constant} > 0$. For a simple model assume $I = S$. Then

$$g(dy/dt) - sy = 0.$$

Hence

$$y = ce^{(s/g)t}.$$

For a more refined model assume

$$y(t) = C(t) + I(t) + A(t),$$

where

$$C(t) = \underline{\text{rate of consumption}}$$

$$= cy(t), \quad 0 < c = \text{constant} < 1,$$

$$A(t) = \underline{\text{autonomous investment rate}}.$$

The function $A(t)$ is independent of $y(t)$ (e.g., Social Security). Now assume $I = g(dy/dt)$. Then

$$y = cy + g(dy/dt) + A(t)$$

($A(t)$ is like an input function). Therefore

$$g(dy/dt) - (1 - c)y = -A(t),$$

which is unstable for $0 < c < 1$. Hence the autonomous investment rate, $A(t)$, is not very important (at least if it is bounded) in the long run, since the exponential growth term, $\exp(t[1 - c]/g)$, will dominate. Also observe that s in the "simple model" corresponds to $1 - c$ in the "refined model." Typical values would be $s = 0.5$, $c = 0.95$.

Phillips model. Replace $I = g(dy/dt)$ by

$$\alpha(dI/dt) + I = g(dy/dt),$$

thereby introducing a lag, and replace $y = C + I + A$ (total demand) by

$$\beta(dy/dt) + y = C + I + A,$$

where we assume as before that $C = cy$ and $1 - c = s$. (Note. The quantity $Z = C + I + A$ is called the total demand). One can then obtain a system of differential equations in the unknown functions y and I of the form

$$\frac{d}{dt}\begin{bmatrix} I \\ y \end{bmatrix} = Q\begin{bmatrix} I \\ y \end{bmatrix} + \vec{F}(t).$$

249

Phillips stabilization policy. In the preceding model replace C + I + A by C + I + A + G. The planned control function G = G(t) satisfies a differential equation of the form

$$\gamma(dG/dt) + G = \overline{G},$$

where $\overline{G} = \overline{G}(t)$ is a computed stabilizing term of the form

$$\overline{G}(t) = -f_p y - f_i \int_0^t y\, dt - f_d(dy/dt).$$

Here f_p, f_i, and f_d are feedback constants. One can then obtain a system of 4 first-order equations:

$$\alpha(dI/dt) + I = g(dy/dt),$$

$$\beta(dy/dt) + y = C + I + A + G,$$

$$\gamma(dG/dt) + G = -f_p y - f_i u - f_d(dy/dt),$$

$$(du/dt) = y, \qquad u(0) = 0.$$

Next one tries to adjust the constants so that the solutions--at least y--are stable. Suppose we assume y, A, and G are constants. Then we get $I \equiv 0$ and

$$y(1 - c) = G + A$$

from the second equation (recall C = cy), and then $G = -f_p y$ from the third equation. Hence

$$y = A/(1 - c + f_p).$$

Samuelson's model for market fluctuations. Let

$I(t)$ = rate of investment,

$K(t)$ = total capital,

$\overline{K}$ = equilibrium value,

(apparently, this is a short-term theory),

$$k(t) = K(t) - \overline{K}.$$

First assume

$$I(t) = -m(K(t) - \overline{K}),$$

$$dK/dt = I(t).$$

Then $dk/dt = -mk$. Therefore $k(t) = k_o e^{-mt}$, which goes to 0; hence the system is stable.

Next assume

$$dI/dt = -m(K(t) - \overline{K}),$$

$$dK/dt = I(t).$$

Then we get $d^2k/dt^2 = -mk$, whose solutions oscillate for $m > 0$.

Goodwin model. Let

K_d = desired capital,

I_{max} = maximum allowed investment,

$I = I_{max}$ for $K < K_d$,

I_{min} = minimum investment level < 0,

(I_{min} corresponds to nonreplacement of productive facilities as they wear out),

$I = 0$ for $K = K_d$.

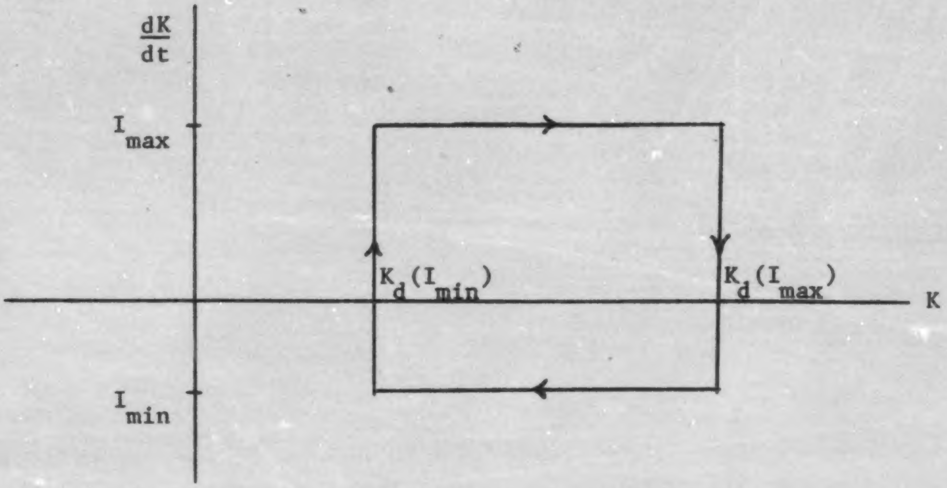
Assume $K_d = ky = k(C + I) = k(ay + b + [dK/dt])$.
Then

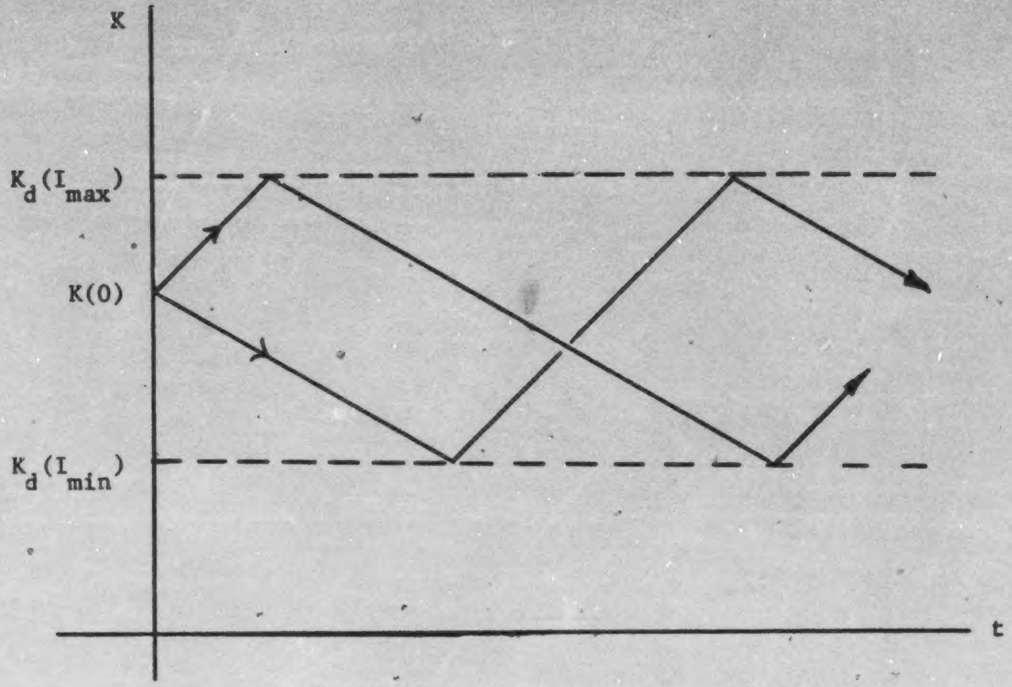
$$y = ay + b + (dK/dt),$$

$$K_d = \frac{kb}{1-a} + \frac{k}{1-a} \cdot \frac{dK}{dt},$$

$$\begin{aligned} \frac{dK}{dt} &= I_{max} \quad \text{for } K_d > K \\ &= 0 \quad\quad \text{for } K_d = K \\ &= I_{min} \quad \text{for } K_d < K . \end{aligned}$$

Graphically:





For a given initial value K(0), $K_d(I_{min}) < K(0) < K_d(I_{max})$, there are two solutions, depending on whether the economy is rising or falling at time $t = 0$. Although this model is mathematically unsatisfactory, economists consider it important.

Lecture 7: Applications of Differential Equations in Biology

References:

Bailey [3] and Homeostasis and Feedback Mechanisms [34].

There is a broad category of applications of mathematics to growth problems in biology. There are also many applications to physical phenomena--mechanics, heat transfer, fluid mechanics, elasticity, electric current, diffusion, etc.--in living organisms. Generally, biological problems are extremely complex, making it very difficult to apply mathematics in a meaningful way.

The predator-prey problem. We assume a single predator population feeding on a single prey population. Let

n_1 = "size" of prey population,

n_2 = "size" of predator population.

(Usually, but not necessarily, the "size" would be given by a numerical count.) We assume:

$$dn_1/dt = n_1(\beta_1 - \mu_1 n_2),\ n_1 \geq 0,$$

$$dn_2/dt = n_2(\beta_2 n_1 - \mu_2),\ n_2 \geq 0,$$

where β_1, β_2, μ_1, μ_2 are all nonnegative constants. These equations have the right characteristics, showing what appears to be reasonable qualitative behavior.

One can write

$$\frac{dn_2}{dn_1} = \frac{n_2(\beta_2 n_1 - \mu_2)}{n_1(\beta_1 - \mu_1 n_2)},$$

separate variables, and integrate, but the resulting equations are difficult to interpret. The equilibrium points are $(n_1,n_2) = (0,0)$ and $(n_1,n_2) = (\mu_2/\beta_2,\beta_1/\mu_1)$. The first of these is a saddle point. Translating to the second equilibrium point, we get:

$$dx/dt = -(\mu_1\mu_2/\beta_2)y - \mu_1 xy$$

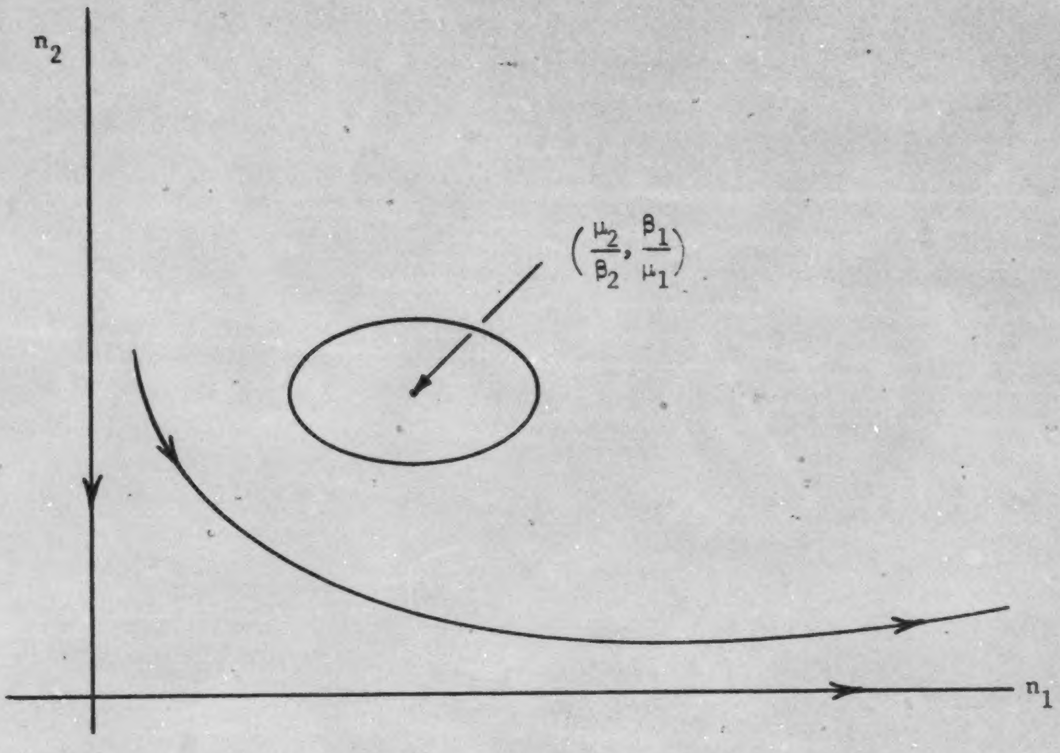
$$dy/dt = (\beta_1\beta_2/\mu_1)x + \beta_2 xy.$$

Linearizing:

$$dx/dt = -(\mu_1\mu_2/\beta_2)y$$

$$dy/dt = (\beta_1\beta_2/\mu_1)x,$$

we get elliptical orbits about the equilibrium point. Hence it is a center.



A path near the axes is likely to result in extinction of the prey due to stochastic variations.

The following values for the parameters were determined from observations of paramecia feeding on yeast cells: $\beta_1 \sim 1$, $\mu_1 \sim 0.01$, $\mu_2 \sim 0.45$, $\beta_2 \sim 3 \times 10^{-8}$, equilibrium $\sim (1.5 \times 10^7, 100)$, period $\sim 2\pi/\sqrt{\mu_2\beta_1}$.

Epidemics. For a simple model describing the spread of an epidemic, let

α = total population

x = number of susceptibles which are not infected.

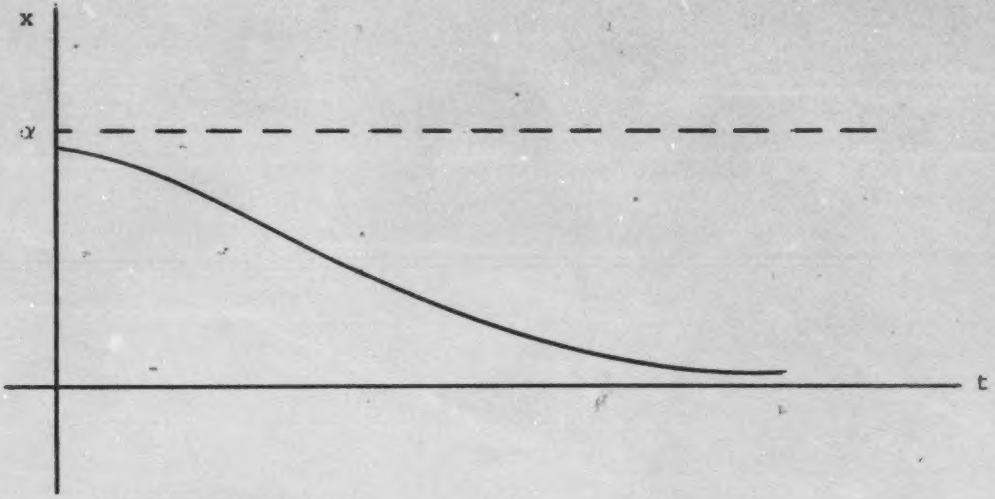
Then $\alpha - x$ = number infected.

Assume

$$dx/dt = -\beta x(\alpha - x),$$

where β is a positive constant.

We indicate the behavior of the solution graphically:



This model is similar to the Law of Mass Action in chemistry.

For a more refined model, let α and x denote the same quantities as before and let

y = number which are infected

z = number removed from activity (i.e., noncarriers which are isolated, dead, or recovered and immune).

Assume

$$dx/dt = -\beta xy$$

$$dy/dt = y(\beta x - \gamma)$$

$$dz/dt = \gamma y,$$

where β and γ are positive constants. Then

$$d(x + y + z)/dt = 0,$$

since $x + y + z = \alpha$.

[Note. Substituting from the first two differential equations assumed in the model:

$$\begin{aligned} dz/dt &= d(\alpha - x - y)/dt \\ &= -(dx/dt) - (dy/dt) \\ &= \beta xy - y(\beta x - \gamma) \\ &= + \gamma y \end{aligned}$$

Hence the three differential equations given in the model are dependent and consistent.]

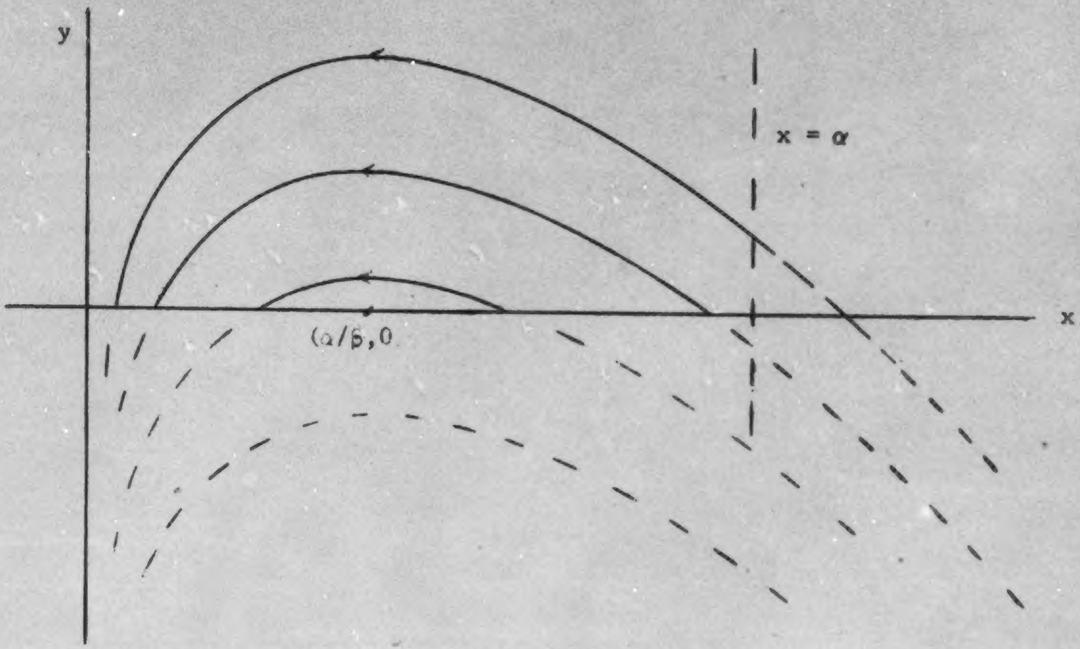
Eliminating z and t:

$$\frac{dy}{dx} = \frac{y(\beta x - \gamma)}{-\beta xy} = -1 + \frac{\gamma}{\beta x}$$

Therefore

$$y = -x + (\gamma/\beta) \log x + C.$$

There is an equilibrium point $(x,y) = (\gamma/\beta, 0)$. The general shape of the graphs in the xy-plane is indicated on the following page.



One observes that x goes from a maximum to a minimum value in a finite time; as x approaches its minimum value (which is always positive), y goes to zero. (Notice that x and y are also bounded above by α, which may be less than, equal to, or greater than γ/β.) [Reference: Bailey [3], p. 190.]

Feedback control in biology. [Various articles from Homeostasis and Feedback Mechanisms [34] are surveyed in the remainder of this lecture.]

K. E. Machin, "Feedback theory." Transfer functions are used to obtain a system of simultaneous linear differential equations with polynomial coefficients:

$$p_1(D)x + p_2(D)y = F_1(t)$$

$$p_3(D)x + p_3(D)y = F_2(t)$$

One can apply differential operations to eliminate y [See Perlis [24].], obtaining (for $F_2 \equiv 0$):

$$Q_1(D)x = Q_2(D)\ F_1(t)$$

Try $F_1(t) = e^{st}$, obtaining

$$x = [Q_2(s)/Q_1(s)]e^{st}.$$

Then $Y(s) = Q_2(s)/Q_1(s)$ is the transfer function. Also, $Y(i\omega)$ is the frequency response function, and

$$|Y(i\omega)| = \text{amplification factor},$$

$$\arg Y(i\omega) = \text{phase shift}.$$

One can obtain a similar result using the Laplace transform, $\mathcal{L}$:

$$\mathcal{L}[x] = Y(s)\,\mathcal{L}[F_1(t)]$$

$$Y(s) = \mathcal{L}[W(r)]$$

If $F_1(t)$ is the Dirac delta function, then

$$\mathcal{L}[F_1] \equiv 1,$$

$$x = W(t).$$

Machin discusses the different kinds of transfer functions which appear for different inputs.

D. H. Fender, "Eye motions." This article studies the movement of the eye in following a moving object. It also considers the mechanics of the eye and the eye as an adaptive system--"We are dealing with an adaptive system which is able to change its own organization."

Goodwin, "A statistical mechanics of temporal organization in cells." Statistical mechanics was developed in the 19th century by Willard Gibbs to explain thermodynamics from the molecular model of a gas--the Kinetic Theory of Gases. One has a system of differential equations

$$(dx_i/dt) = f_i(x_1, \ldots, x_n), \quad i = 1, \ldots, n,$$

where each x_i is a position vector and n is very large--of the order 10^{23}. One looks for "average behavior," obtaining an energy integral which can be related to the average kinetic energy of the various molecules.

Goodwin adopts this point of view in studying the development of organisms from "genetic codes." Let

X = RNA messenger

Y_i = protein.

These satisfy the differential equations

$$\frac{dx_i}{dt} = \frac{a_i}{A_i + k_i Y_i} - b_i$$

$$\frac{dY_i}{dt} = \alpha_i X_i - \beta_i .$$

By linearizing, one can predict oscillations about an equilibrium point. Eliminating t, one obtains the integral

$$G_i(X_i,Y_i) = (\alpha_i/2)\, X_i^2 - \beta_i X_i + b_i Y_i$$

$$- (a_i/k_i) \log (A_i + k_i Y_i) = \text{constant},$$

which is like an energy integral. Adding, we get

$$G(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \sum_{i=1}^{n} G_i(X_i,Y_i).$$

Goodwin also obtains a "talandic" temperature, which is a measure of how fast the process is going.

DIFFERENTIAL EQUATIONS

Lecture 8: Remarks on Centers, Boundary Value Problems

Remarks on centers. A center is generally not structurally stable. For example, consider

$$dx/dt = -y + P(x,y)$$

$$dy/dt = x + Q(x,y), \qquad (*)$$

where P and Q are polynomials (or power series) of order 2 or higher (i.e., all terms in P and Q are of degree 2 or higher). The linear terms indicate a center at (0,0), but this center can be spoiled by P, Q. Considering polar coordinates r,θ, we obtain

$$rr' = xx' + yy' = 0$$

$$r^2\theta' = -yx' + xy' = r^2$$

for the linear terms. Assuming all of the terms on the right side in (*), we obtain

$$rr' = xP + yQ$$

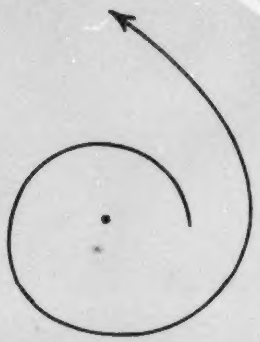
$$r^2\theta' = r^2 - yP + xQ.$$

Hence, $\theta' > 0$ near (0,0) for the functions x(t), y(t) described by equations (*), but r' can go back and forth from positive to negative.

Suppose $P = x^3$ and $Q = y^3$. Then

$$rr' = x^4 + y^4.$$

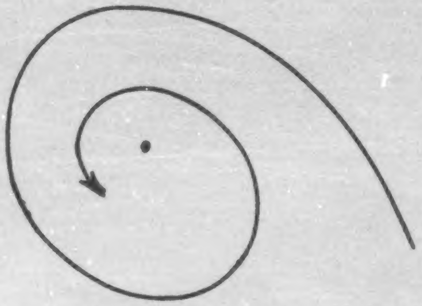
Hence, $r' > 0$ for $r \neq 0$, and a solution will spiral away from the origin.



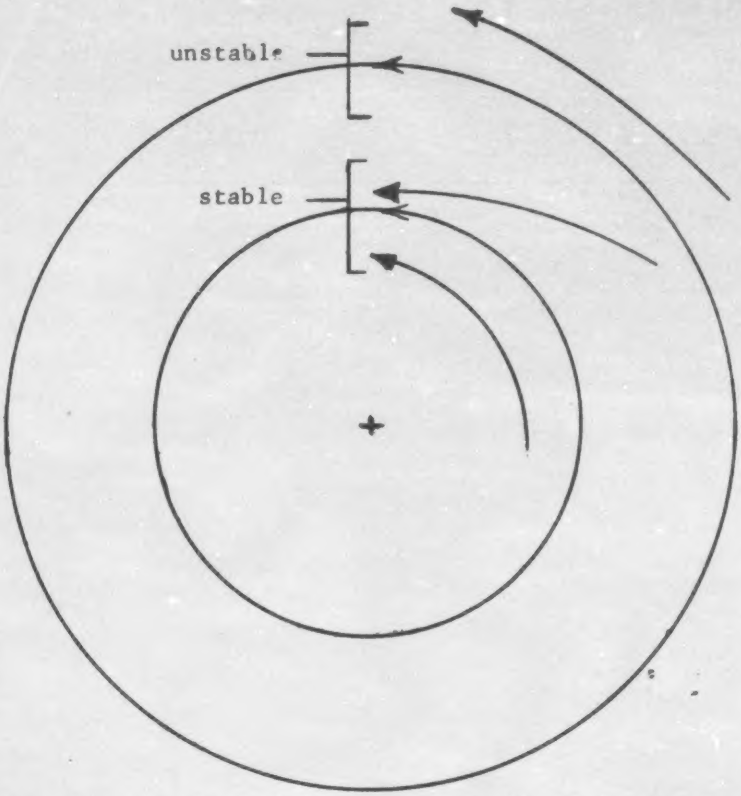
If $P = -x^3$ and $Q = -y^3$, then

$$rr' = -x^4 - y^4.$$

Hence $r' < 0$ for $r \neq 0$, and a solution will spiral into the origin.

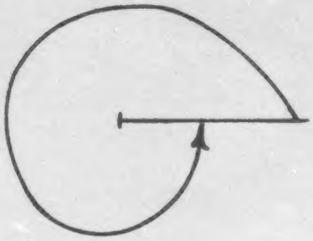


Suppose $rr' = xP + yQ = g(r)$, where $g(r)$ cscillates but goes to zero at $r = 0$, e.g., $g(r) = r^2 \sin r^2$. Then we get alternately stable and unstable limit cycles.



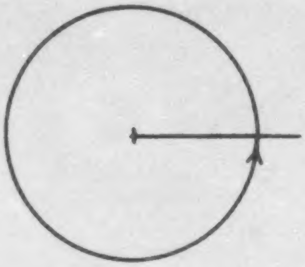
If the differential equation $dy/dx = (-y + P)/(x + Q)$, obtained by eliminating t from (*), has a first integral $\phi(x,y) = c$ such that $\phi(x,y)$ is continuous for all x,y and is not identically constant on any nonempty open set in the xy-plane, then (by continuity arguments) there cannot exist a limit cycle or a focus.

Now suppose $d\theta/dt > 0$ and $|dr/dt| < \epsilon$ in some neighborhood of (0,0). Then a solution winds around the origin. Consider a solution starting at (r_o,θ_o) and making one complete revolution, ending at $(r_1, \theta_o + 2\pi)$. There are three cases: $r_1 < r_o$, $r_1 = r_o$, or $r_1 > r_o$.



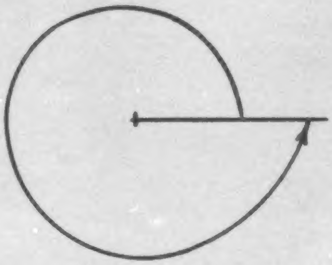
Case 1

$r_1 < r_o$



Case 2

$r_1 = r_o$



Case 3

$r_1 > r_o$

In cases 1 and 3 the solutions, if continued, would spiral to a limit cycle or a focus. Hence, if our equation has a first integral satisfying the conditions mentioned above, then only case 2 can obtain: there is a center.

Boundary value problem. For a second-order differential equation

$$d^2y/dx^2 = f(x, y, dy/dx)$$

we can apply two arbitrary conditions (of a suitable type)--say $y(x_o) = y_o$, $y(x_1) = y_1$, resulting in a boundary value problem (BVP). We can interpret this problem as: Given the family of solutions through (x_o,y_o), find the members of this family which pass through (x_1,y_1).

Example. Great circles on a sphere. For (x_o,y_o) and (x_1,y_1) close together, there are two solutions--the two arcs of the unique great circle through the two points. If (x_o,y_o) and (x_1,y_1) are diametrically opposite, there are infinitely many great circles through them.

Example. Given $y'' + y = 0$.

(a) For the boundary conditions $y(0) = 0$, $y(\pi) = 0$, we have infinitely many solutions: $y = C \sin x$ for every C.

(b) For the boundary conditions $y(0) = 0$, $y(\pi) = 1$, there are no solutions.

(c) For the boundary conditions $y(0) = 1$, $y(\pi/4) = \sqrt{2}/2$, there is a unique solution: $y = \cos x$.

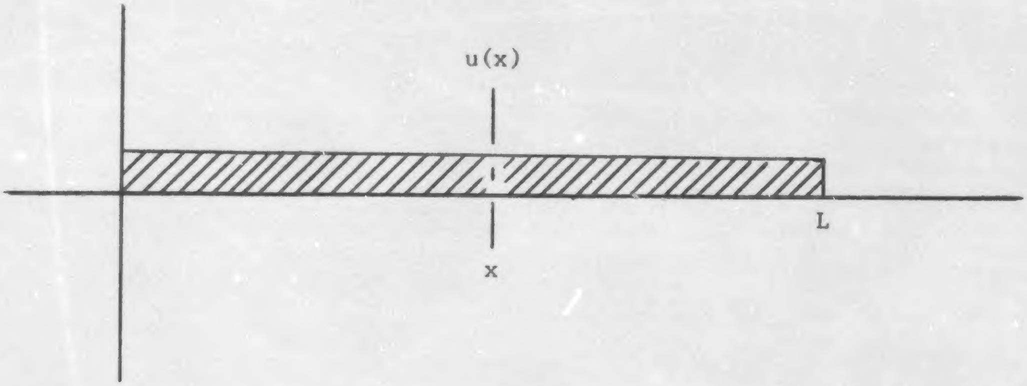
Theorem of Picard. For any differential equation of second order satisfying appropriate continuity conditions, for each (x_o, y_o) there exists $\epsilon > 0$ such that there exists a unique solution through (x_o, y_o) and (x_1, y_1) if $\sqrt{(x_1 - x_o)^2 + (y_1 - y_o)^2} < \epsilon$.

Example. Vibrating string of length π. For a fixed t, let the displacement $y = A(x)$. Then one can show

$$a^2 A''(x) + \lambda^2 A(x) = 0, \quad A(0) = A(L) = 0,$$

where a is a constant dependent on the physical characteristics of the string. (If t is not fixed, one obtains solutions $y(x,t) = A(x) \sin \lambda t$.) Then $y = \sin \lambda_n x$, $\lambda_n = an$, $n = 1, 2, 3, \ldots$, satisfies the boundary value problem. Any linear combination (superposition) of the solutions $\sin \lambda_n x$ is also a solution. A solution of the form $k \sin \lambda_1 x$ (i.e., $n = 1$) is called the fundamental tone and a solution $k \sin \lambda_2 x$ ($n = 2$) is called the first overtone.

Example. One-dimensional heat conduction. Let $u(x)$ be the temperature at x in a long thin bar of length L, $0 \leq x \leq L$.



We assume the temperature is at equilibrium (i.e., is independent of time). Suppose there are internal sources of heat and that the ends of the bar are

263

kept at specified temperatures. Then u must satisfy the boundary value problem

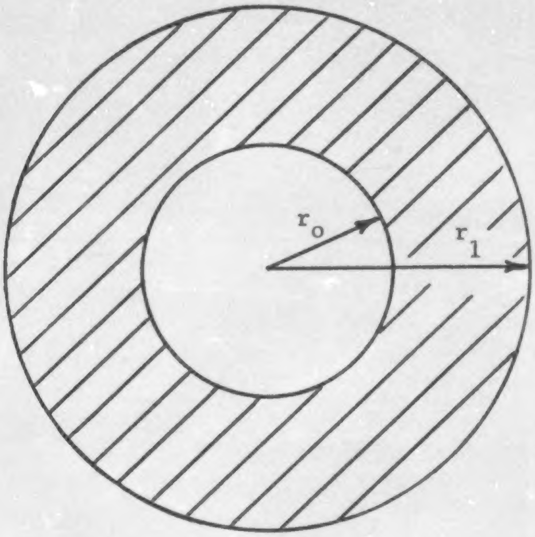
$$-K^2 \, d^2u/dx^2 = F(x), \quad u(0) = u_o, \quad u(L) = u_1,$$

where F(x) is a given input function and u_o and u_1 are the specified temperatures. This boundary value problem has a unique solution of the form

$$u = Ax + B - \int_0^x F(\xi)/K^2 \, d\xi,$$

with A and B adjusted to fit the boundary conditions.

Example. Heat conduction in an annulus--equilibrium, temperature $u = u(r)$, no internal source of heat.



The boundary value problem is

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} = 0, \quad u(r_o) = u_o, \quad u(r_1) = u_1.$$

One can rewrite the differential equation as

$$\frac{d}{dr} \left(r \frac{du}{dr}\right) = 0,$$

obtaining a unique solution of the form

$$u = A \log r + B,$$

with A and B chosen to fit the boundary conditions.

Example. Elasticity of a loaded annulus. The stress function W(r) must satisfy the differential equation

$$W'''' + (2/r)\, W''' - (1/r^2)\, W'' + (1/r^3)\, W' = 0,$$

with W and W' given at r_o and r_1. [See figure above for heat conduction in an annular ring.]

For examples of loaded beams, sagging chains, etc., see Scarborough [28].

Example. Vibrating string of nonuniform density. One gets a boundary value problem of the form

$$K^2\, A''(x) + \rho(x)\, \lambda^2\, A(x) = 0, \quad A(0) = A(L) = 0.$$

One gets a sequence of eigenvalues $\lambda_1, \lambda_2, \ldots$ with corresponding eigenfunctions similar to sine functions. See Bôcher [7].

Example. Vibrating circular membrane (drumhead). Assuming radial symmetry and a radius of 1, i.e., $0 \leq r \leq 1$, the displacement, $z = CR(r)$, must satisfy the differential equation

$$(rR')' + \lambda r R = 0,$$

with the boundary conditions $R(1) = 0$, R continuous at $r = 0$. One obtains as solutions

$$R(r) = J_o(\sqrt{\lambda_n}\, r),$$

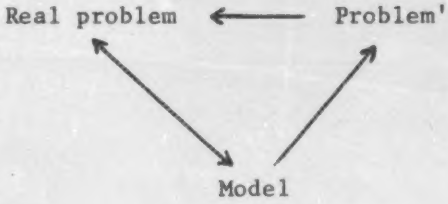
where $\lambda = \lambda_n$ is the n^{th} zero of the Bessel function $J_o(r)$.

DIFFERENTIAL EQUATIONS

Lecture 9: Simple Pendulum and Two-Particle Problems

Differential equations which are encountered in industry may be unsolvable in the sense that they cannot be reduced to quadratures, but one may still be able to obtain a usable "solution" by numerical analysis if the problem is properly stated. In solving a differential equation by incremental methods, one may encounter the "bifurcation problem"--that is, the form of the solution separates at some point. One can show that you will get a satisfactory solution of specified accuracy if your computational methods are sufficiently refined. However it is difficult to make a priori estimates of how refined the method should be.

Applications in mathematics involve the modeling process:



The whole process is an experimental science (or art?).

What special contribution can the mathematician make that the computer programmer cannot make? Unfortunately, a kind of Gresham's Law seems to be in effect: concise, meaningful mathematical analyses are being replaced by copious print-outs of relatively meaningless numerical data. Suppose the modern high-speed computer had been available in the 17th century and the sine function were known only in terms of its power series. Would the programmers have yet discovered that the sine function is periodic?

Let us consider the kind of special contribution that a mathematician can make to the solution of differential equations of the form

$$(d^2x/dt^2) + f(x) = 0 .$$

The model for the problem of undamped motion by a weight suspended by an elastic spring is a linear differential equation

$$x'' + k^2x = 0,$$

and one can write explicit solutions in terms of elementary functions. For a simple pendulum, on the other hand, one gets a nonlinear equation

$$x'' + k^2 \sin x = 0.$$

Expanding sin x in power series, one gets

$$x'' + k^2x = (\text{sum of higher powers in } x).$$

Let us look at the simplest equation of this type

$$x'' + k^2x = ax^2. \qquad (*)$$

[A general method, not a second approximation to the pendulum equation, is being presented here. To improve on the linearization of the pendulum equation, one would have to go to at least a third-order term, i.e., $x'' + k^2x = (k^2/6)x^3$, since the power series of sin x does not have an x^2 term.] By the method of variation of parameters, one can write the solution to (*) as

$$x(t) = C_o \sin (kt - \phi_o) + (a/k) \int_o^t \sin [k(t - \tau)]\, x^2(\tau)\, d\tau.$$

Observe that the first term on the right is the solution to the linearized equation, $x'' + k^2x = 0$. One can try to estimate the second term.

Next we look at the equation

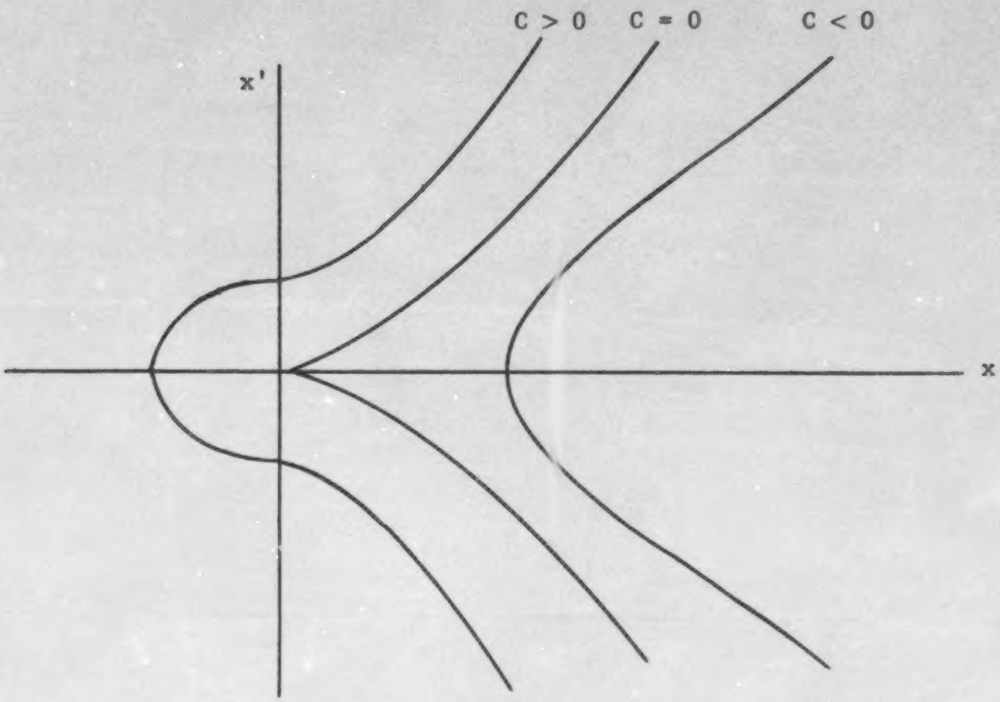
$$x'' = ax^2.$$

This is nonlinear, but we can estimate its behavior for large x by comparing it with $x' = ax$, which suggests that $x = a + bt$ is not a good estimate. For a more refined analysis:

$$
\begin{aligned}
& x'' = ax^2 \\
\Longrightarrow \quad & x''x' = ax^2x' \\
\Longrightarrow \quad & \frac{d}{dt}\left[\frac{(x')^2}{2}\right] = \frac{d}{dt}\left[\frac{ax^3}{3}\right] \\
\Longrightarrow \quad & (x')^2 = \tfrac{2}{3}ax^3 + C.
\end{aligned}
$$

Hence, if x is large, then x' is large. Also, if x is a large negative number, then the right side is negative, so x' is not defined.

Graphically:



Since x has a minimum value for a given value of C, the solutions do not behave like a + bt.

We could solve the differential equation $(x')^2 = (2a/3)x^3 + C$ as accurately as we want by first reducing it to quadratures and then evaluating the integral numerically, but a programmer would solve it by increments. The resulting print-outs could be compared with our phase diagram.

Returning to the pendulum problem:

$$x'' + k^2 \sin x = 0$$

$$\Longrightarrow \quad x'x'' + k^2 x' \sin x = 0$$

$$\Longrightarrow \quad d[\tfrac{1}{2}(x')^2 - k^2 \cos x]/dt = 0$$

$$\Longrightarrow \quad (x')^2 = C + 2k^2 \cos x.$$

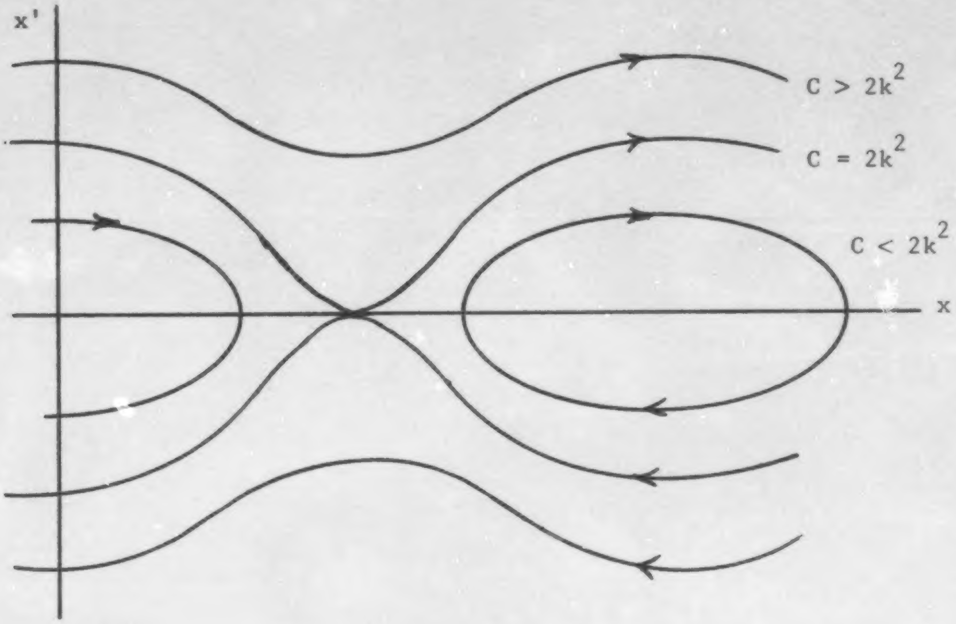
Observe: $\frac{1}{2}(x')^2 - k^2 \cos x$ = kinetic energy + potential energy. Hence our differential equation asserts that the total energy is constant.

Case 1: $C > 2k^2$. Then $C + 2k^2$ cox $x > 0$ for all x. In this case we have an escape situation. The pendulum has enough energy so that it will go around and around. It does not oscillate.

Case 2: $C = 2k^2$. Then $C + 2k^2 \cos x \geq 0$ and is 0 for $x = \pi$, 3π, etc. This is the limiting case between case 1 above and case 3 below. If the pendulum is put in motion with energy $C/2 = k^2$, then it will continue to move in the same direction, reaching the highest possible point while its velocity goes to zero in infinite time.

Case 3: $C < 2k^2$. In this case there are values of x such that $C + 2k^2 \cos x < 0$. For these values of x, the curve is not defined. This is a "capture situation."

Graphically:



In general, for the differential equation $x'' + f(x) = 0$ we have

$$x'[x'' + f(x)] = 0.$$

Hence

$$\frac{d}{dt}\left[\frac{(x')^2}{2} + \int_{x_o}^{x} f(\tau)\, d\tau\right] = 0.$$

Therefore

$$\tfrac{1}{2}(x')^2 + \int_{x_0}^{x} f(\tau)\, d\tau = C.$$

Physically, we may interpret $x'' + f(x)$ as the net force acting on a particle, $\frac{1}{2}(x')^2$ as the kinetic energy of the particle, and $\int_{x_0}^{x} f(\tau)\, d\tau$ as the potential energy of the particle. The original second-order equation, which is a statement about forces, can be reduced to a first-order equation which is a statement about energy. The condition--energy = constant--determines a family of curves in the x,x'-plane which describe all possible motions of the particle consistent with the differential equation.

A mathematician can make a genuine contribution to an "industrial" problem by this type of analysis, i.e., determining the general behavior of solutions of a differential equation without explicitly solving the equation in terms of familiar functions. Most engineers can only handle a problem by linearizing it.

The two-particle problem. We assume a large body at the origin and a small particle whose position is given by polar coordinates (r,θ). We assume that the mass of the second object (henceforth referred to as "the particle") is so small that it does not affect the motion of the first. Let

$$\underline{r} = r\hat{r}$$

be the position vector from the origin to the particle. Here $\hat{r}$, $\hat{\theta}$, $\hat{x}$, $\hat{y}$ are unit vectors. Then:

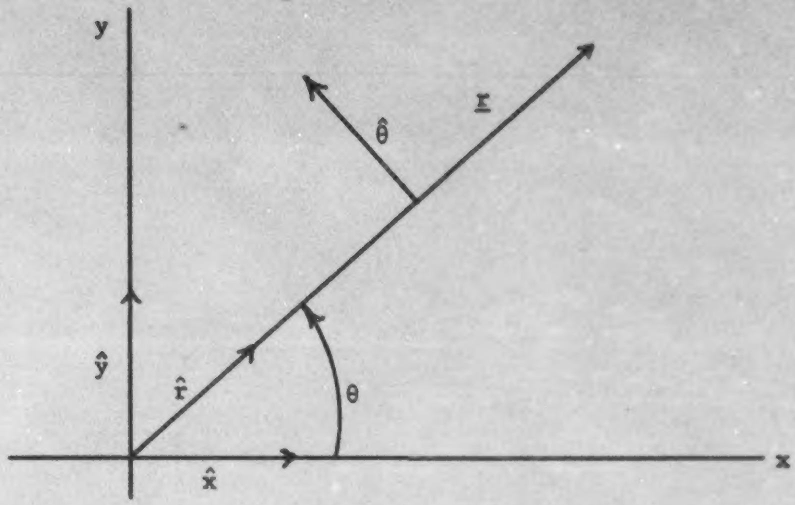
$$\hat{r} = \cos\theta\,\hat{x} + \sin\theta\,\hat{y}$$

$$\hat{\theta} = -\sin\theta\,\hat{x} + \cos\theta\,\hat{y}$$

$$d\hat{r}/d\theta = \hat{\theta} \qquad d\hat{\theta}/d\theta = -\hat{r}$$

$$d\hat{r}/dt = (d\hat{r}/d\theta)(d\theta/dt) = \theta'\hat{\theta}$$

$$d\hat{\theta}/dt = (d\hat{\theta}/d\theta)(d\theta/dt) = -\theta'\hat{r}$$



The equation of motion which we must solve is

$$\underline{r}'' = f(r)\,\hat{r}, \tag{1}$$

where f(r) is the magnitude of the central force on the particle. From $\underline{r} = r\hat{r}$ we get

$$\underline{r}' = r'\hat{r} + r\,(d\hat{r}/dt)$$

$$\underline{r}' = r'\hat{r} + r\theta'\hat{\theta}. \tag{2}$$

Hence

$$\underline{r}'' = r''\hat{r} + 2r'\theta'\hat{\theta} + r\theta''\hat{\theta} + r(\theta')^2(-\hat{r}).$$

Equating components in the equation of motion, we get

$$r'' - r(\theta')^2 = f(r) \tag{3}$$

$$2r'\theta' + r\theta'' = 0. \tag{4}$$

We define the angular momentum $\bar{h}$ to be $r^2\theta'$. (Our unit of mass is such that the particle is of mass 1.) Then

$$d\bar{h}/dt = 2rr'\theta' + r^2\theta''$$

$$= r(2r'\theta' + r\theta'').$$

Hence, from equation (4),

$$d\bar{h}/dt = r \cdot 0 = 0.$$

Therefore equation (4) implies: The angular momentum is conserved.

Multiplying equation (3) by r', we get

$$\begin{aligned} 0 &= r'r'' - r'r(\theta')^2 - f(r)r' \\ &= r'r'' - (r'/r^3)[r^2(\theta')]^2 - f(r)r' \\ &= r'r'' - (r'/r^3)\bar{h}^2 - f(r)r' \\ &= \frac{d}{dt}\left[\frac{(r')^2}{2} + \frac{\bar{h}^2}{r^2} - \int_{r_o}^{r} f(p)\,dp \right]. \end{aligned}$$

Therefore

$$\tfrac{1}{2}[(r')^2 + (\bar{h}/r)^2] - \int_{r_o}^{r} f(p)\,dp = C. \tag{5}$$

From equation (2), the square of the speed of the particle is

$$\begin{aligned} |\underline{r}'|^2 &= (r')^2 + (r\theta')^2 \\ &= (r')^2 + (\bar{h}/r)^2. \end{aligned}$$

Hence, $\frac{1}{2}[(r')^2 + (\bar{h}/r)^2]$ = kinetic energy. Since $\int_{r_o}^{r} f(p)\,dp$ = work done by force of magnitude $f(p)$ in moving particle from r_o to r, i.e., $-\int_{r_o}^{r} f(p)\,dp$ = potential energy, equation (5) asserts: The total energy of the particle is conserved. If we assume the inverse square law for the force of attraction on the particle, $f(r) = -c/r^2$, then

$$-\int_{r_o}^{r} f(p)\,dp = c\int_{r_o}^{r} p^{-2}\,dp = -c(r^{-1} - r_o^{-1}).$$

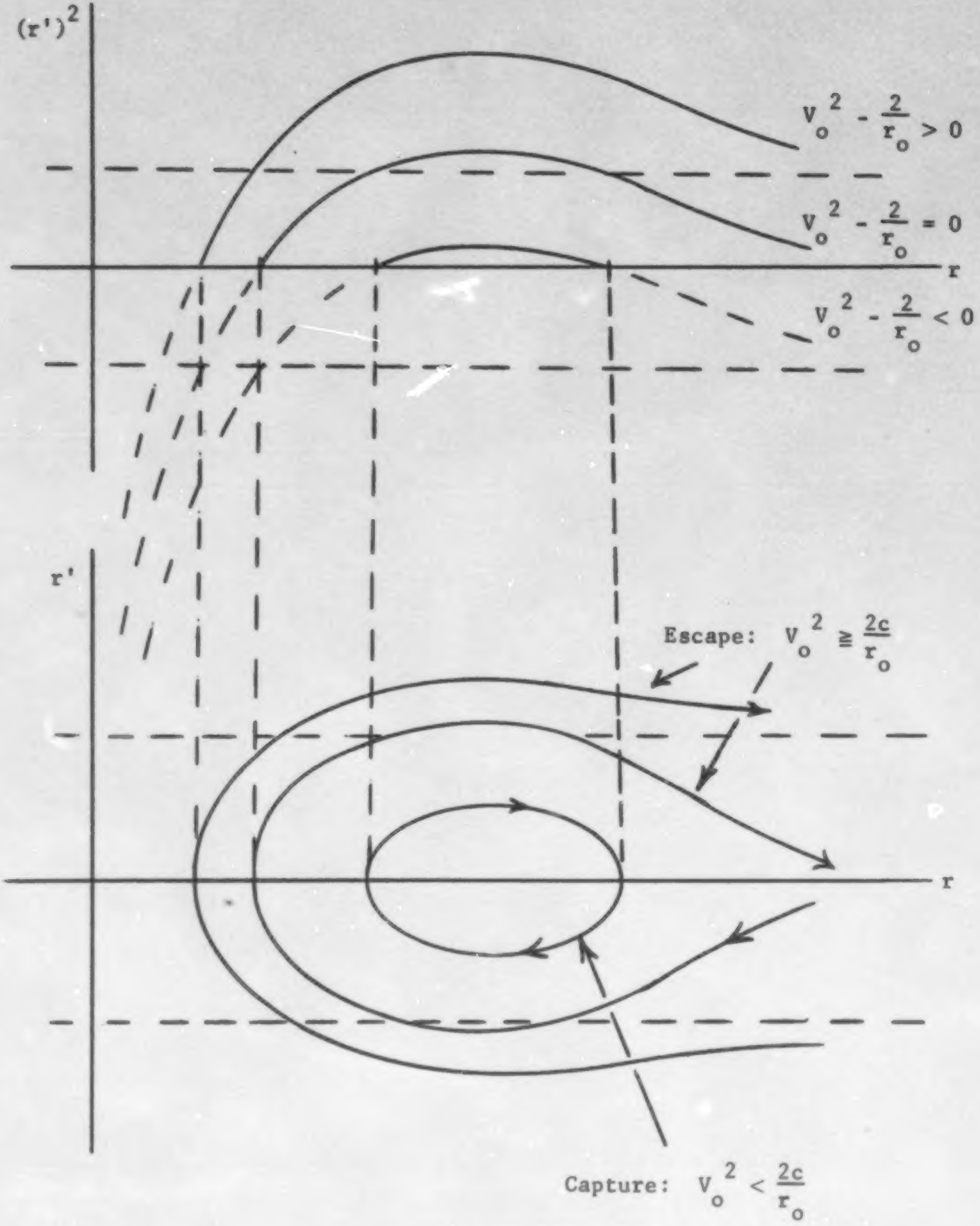
Substituting in equation (5) and setting $C = \frac{1}{2}V_o^2$, we get

$$(r')^2 + \frac{\bar{h}^2}{r^2} - \frac{2c}{r} = V_o^2 - \frac{2c}{r_o},$$

or

$$(r')^2 = \left[\frac{2c}{r} - \frac{\bar{h}^2}{r^2}\right] + \left[V_o^2 - \frac{2c}{r_o}\right].$$

Graphing $(r')^2$ vs. r and r' vs. r:


$(r')^2$
$V_o^2 - \frac{2}{r_o} > 0$
$V_o^2 - \frac{2}{r_o} = 0$
r
$V_o^2 - \frac{2}{r_o} < 0$
r'
Escape: $V_o^2 \geq \frac{2c}{r_o}$
r
Capture: $V_o^2 < \frac{2c}{r_o}$

We see that the particle is captured in an orbit if the energy V_o^2 is less than $2c/r_o$. If the energy is greater than or equal to $2c/r_o$, the particle escapes.

Since $\bar{h} = r^2\theta'$ is a constant,

$$d(r^{-1})/d\theta = -r^{-2}(r'/\theta') = -r'/(r^2\theta') = -r'/\bar{h}.$$

Therefore

$$\frac{d^2}{d\theta^2}\left(\frac{1}{r}\right) = -\frac{r''}{\bar{h}} \cdot \frac{dt}{d\theta} = -\frac{r''}{\bar{h}\theta'} .$$

From equation (3) and the inverse square law, $f(r) = -c/r^2$, we obtain

$$\frac{r''}{\bar{h}\theta'} = \frac{f(r) + r(\theta')^2}{\bar{h}\theta'} = \frac{-c/r^2 + r(\theta')^2}{\bar{h}\theta'}$$

$$= \frac{-c + r^3(\theta')^2}{\bar{h}r^2\theta'} = \frac{-c}{\bar{h}^2} + \frac{1}{r}\left[\frac{r^4(\theta')^2}{\bar{h}r^2\theta'}\right]$$

$$= -\frac{c}{\bar{h}^2} + \frac{1}{r} .$$

Therefore

$$\frac{d^2}{d\theta^2}\left(\frac{1}{r}\right) + \frac{1}{r} = \frac{c}{\bar{h}^2} .$$

Hence

$$\frac{1}{r} = \frac{c}{\bar{h}^2} + A \sin (\theta + \phi_o).$$

Therefore

$$r = \frac{\bar{h}^2}{c + \bar{h}^2 A \sin (\theta + \phi_o)} .$$

Hence: (Kepler's First Law). The path of the particle is a conic section.

Let $\alpha(\theta)$ be the area swept out by the radius vector to the particle as the radius vector moves from $\theta = 0$ to θ. Then $(d\alpha/d\theta) = \frac{1}{2}r^2$ and

$$\frac{d\alpha}{dt} = \frac{d\alpha}{d\theta}\frac{d\theta}{dt} = \frac{1}{2}r^2\theta' = \frac{1}{2}\bar{h} = \text{constant}.$$

Hence: (Kepler's Second Law). The radius vector to the particle sweeps out area at a constant rate.

One can also derive Kepler's Third Law from these equations.

Problem. Carry out a similar analysis (as far as you can) for a central force $f(r) = -c/r^3$.

DIFFERENTIAL EQUATIONS

Lecture 10: Predator-Prey Model, Van der Pol Equation, Poincaré-Bendixson Theorem

Further analysis of the predator-prey model [introduced in lecture 7 by W. Kaplan]. Consider the differential equations

$$dx/dt = x(1 + y)$$

$$dy/dt = -y(1 + x).$$

The critical points are (0,0) and (-1,-1) in the xy-plane. Linearizing at (0,0) gives the matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, which indicates a saddle point. Linearizing at (-1,-1) gives $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, which indicates a center. If the linearized equation at (-1,-1) has a center, what about the original equation? The general relationship is indicated by the following table:

Linear		Nonlinear
asymptotically stable	⟷	asymptotically stable
unstable	⟷	unstable (for an arbitrarily small neighborhood, there exists a solution which escapes)
center (eigenvalues are purely imaginary)	⟷	no definite conclusion (eigenvalues may have very small positive or negative real part)

We look now at the differential equations in the predator-prey model:

$$dn_1/dt = n_1(1 - n_2)$$

$$dn_2/dt = n_2(n_1 - 1)$$

To study the critical point (1,1), we set $x = n_1 - 1$ and $y = n_2 - 1$, obtaining

$$dx/dt = -y(x + 1)$$

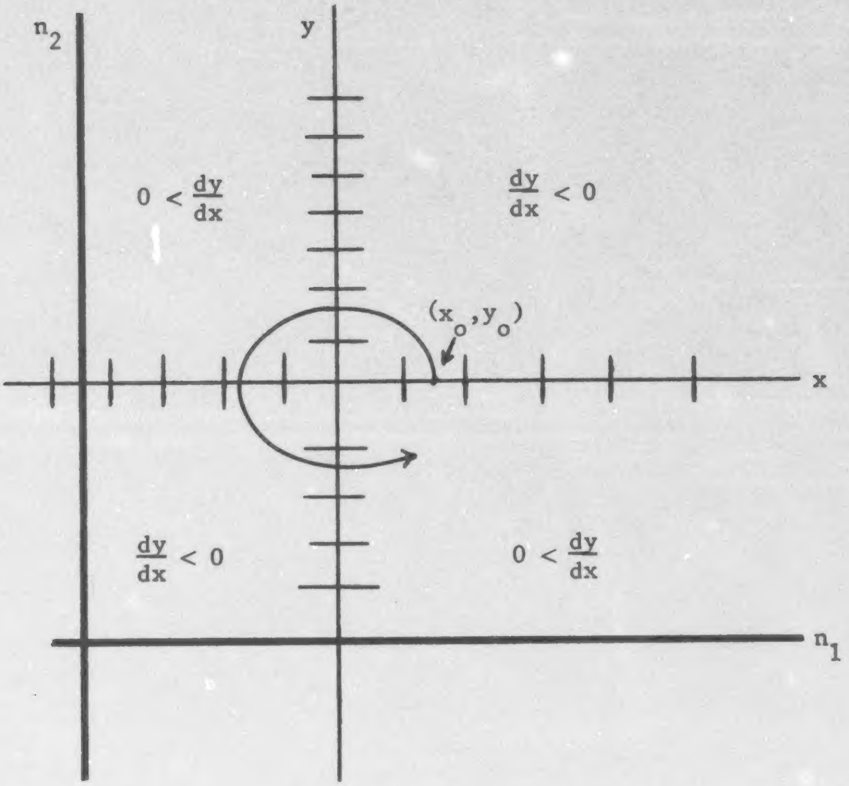
$$dy/dt = x(y + 1).$$

Eliminating t, we get

$$\frac{dy}{dx} = \frac{-x(y + 1)}{y(x + 1)},$$

which has the solution curves

$$x + y - \log (1 + x)(1 + y) = C.$$



Suppose we consider a solution starting at a point (x_o,y_o) near the critical point $(x,y) = (0,0)$. This solution will go around the critical point, but will it come back to (x_o,y_o) or will it spiral in or out? Will we get limit cycles? Kaplan: "No, because of the nicety of the solution." Closure is indicated by the fact that if we expand $x + y - \log (1 + x)(1 + y)$ in a power series, we get

$$x + y - \log (1 + x) - \log (1 + y)$$
$$= x + y - (x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots) - (y - \frac{y^2}{2} + \frac{y^3}{3} - \ldots)$$
$$= \tfrac{1}{2}(x^2 + y^2) + \text{(higher-order terms)}.$$

The Van der Pol equation:

$$x'' + \mu(x^2 - 1)x' + x = 0, \quad \mu \geq 0$$

If $\mu = 0$, we get simple harmonic motion. For $\mu > 0$, if $|x| > 1$ then $\mu(x^2 - 1) > 0$. Hence we get damping--energy is going out of the system. If $|x| < 1$, then $\mu(x^2 - 1) < 1$ and we get anti-damping--energy is being fed into the system.

One can show that for a given value of μ, the Van der Pol equation has exactly one stable limit cycle in the phase plane. If one starts sufficiently near the origin, the solutions wind out. If one starts sufficiently far from the origin, they wind in.

In nature one does not get pure oscillation from a linear differential equation. It is obtained by filtering a solution from a nonlinear differential equation.

Poincaré-Bendixson Theorem. Given the system of differential equations

$$x' = f(x,y), \quad y' = g(x,y).$$

If D is a region which does not contain any critical points and if every solution which starts in D stays there, then there exists at least one periodic solution, and every solution "converges" to a periodic solution.

This is really a topological theorem--not a differential equations theorem. Its proof is too difficult for an elementary course, but one can convince students by having them try to construct counter-examples.

If we apply the Poincaré-Bendixson Theorem to Van der Pol's equation, what is the relationship of the period to that of the free oscillation $\mu = 0$? One can get a good solution for μ very large or μ very small by setting

$$x = x_1 + x_2\mu + x_3\mu^3 + \ldots,$$

substituting, getting a sequence of differential equations, and solving them in turn.

Lecture 11: One-dimensional Eigenvalue Problems

One-dimensional eigenvalue problems. Consider the problem

$$iy' = \lambda y, \quad |y(0)| = |y(1)|. \qquad (*)$$

We rewrite the boundary conditions:

$$y(1) = e^{i\alpha}y(0)$$

Solving the differential equation, we get

$$y = e^{-i\lambda x}.$$

Substituting from the boundary conditions, we get

$$e^{-i\lambda} = e^{i\alpha}.$$

Hence

$$\lambda = \lambda_n = 2\pi n - \alpha, \quad n = 0, 1, 2, \ldots$$

This is mathematically meaningful, but is it physically meaningful?

Consider a particle moving along the real axis with energy E in terms of de Broglie (pre-Schrödinger) quantum mechanics. Then

Ψ = the wave function for the particle.

We interpret

$|\Psi|^2$ = probability density,

i.e.,

$\int_a^b |\Psi|^2\, dx$ = probability that the particle is in the interval (a,b).

The momentum P is replaced by the operator i d/dx, i.e., the momentum of the particle is i $d\Psi/dx$. The Hamiltonian $\frac{1}{2}m^2P^2 + V(x) = E$ becomes

$$-\tfrac{1}{2}m\Psi'' + V(x)\,\Psi = E\Psi,$$

which is of the form

$$y'' + (\lambda - V)y = 0, \qquad (**)$$

i.e., we have a one-dimensional Sturm-Liouville type problem.

Returning to our problem (*):

$$iy' = \lambda y, \quad |y(0)| = |y(1)|$$

If we take $y = \Psi$, then we are placing a constraint on a particle so that the probability of its being at $x = 1$ is the same as the probability of its being at $x = 0$. Our solution to (*) indicates that this can be done for only certain discrete values of the momentum.

Now

$$\int_{-\infty}^{\infty} |\Psi|^2 \, dx = 1,$$

so for any function $y(x) = \text{constant} \cdot \Psi(x)$, $y \to 0$ as $x \to \pm\infty$. Hence (**) becomes a boundary value problem

$$y'' + (\lambda - V)y = 0, \quad y \to 0 \text{ as } x \to \pm\infty. \qquad (***)$$

Example. Harmonic oscillator: $V = x^2$.

The differential equation (***) becomes

$$y'' = (x^2 - \lambda)y.$$

Consider the asymptote in the positive direction. For $y > 0$ and x sufficiently large, $y'' > 0$. Hence, the curve will be concave upward. We conclude that of all the solutions through a given point (x_0, y_0) only one solution of the differential equation will go to zero as $x \to +\infty$. Will the same solution go to zero as $x \to -\infty$?

Suppose we try

$$y(x) = w(x) \exp(-x^2/2),$$

where

$$w(x) = \sum_{n=0}^{\infty} a_n x^n.$$

Then

$$y'(x) = [w' - xw] \exp(-x^2/2)$$

$$y''(x) = [w'' - 2xw' - w + x^2 w] \exp(-x^2/2).$$

Substituting, we obtain

$$w'' - 2xw' + (\lambda - 1)w = 0.$$

Substituting from the power series expansion $w(x) = \Sigma a_n x^n$ we get

$$a_{n+2} = \frac{(2n + 1 - \lambda)}{(n+1)(n+2)} a_n .$$

Therefore $a_{n+2}/a_n \sim 2/n$, which means that the series is like the series for $\exp x^2$. Hence, $y = w \exp(-x^2/2)$ "flunks" unless the series terminates, i.e., $\lambda = 2n + 1$ for some $n = 0, 1, 2, \ldots.$

Let us now look at the case $\lambda = 0$. The differential equation becomes

$$y'' + x^2 y = 0.$$

What happens for large x? Try powers of $t = 1/x$. Then the differential equation becomes

$$\frac{d^2y}{dt^2} + \frac{2}{t}\frac{dy}{dt} + \frac{1}{t^6}y = 0.$$

Try Frobenius' method. We are in trouble, but we go ahead. Next try

$$\begin{aligned} w(x) &= x^{-\frac{1}{2}} \exp(ix^2/2) \\ &= \exp\left[\tfrac{1}{2}(ix^2 - \log x)\right] \end{aligned}$$

(due to Liouville), then

$$dw/dx = \left(ix - \frac{1}{2x}\right) w(x)$$

$$\begin{aligned} \frac{d^2w}{dx^2} &= \left(ix - \frac{1}{2x}\right)^2 w + \left(i + \frac{1}{2x^2}\right) w \\ &= \left(-x^2 - i + \frac{1}{4x^2} + i + \frac{2}{4x^2}\right) w. \end{aligned}$$

Therefore

$$w'' + \left(x^2 - \frac{3}{4x^2}\right) w = 0.$$

Hence, for large x we would expect

$$y(x) \sim w(x) = x^{-\frac{1}{2}} \exp(ix^2/2),$$

where $y'' + x^2 y = 0.$

Now try

$$y(x) = w(x)\, F(x).$$

We expect F to be nearly constant for large x, so we try

$$F(x) = 1 + \frac{F_o}{x} + \frac{F_1}{x^2} + \ldots.$$

But this series cannot converge, for convergence would contradict the Frobenius theory. If one takes only a finite part of the series, then one gets a useful result:

$$[(y/w) - 1]x \to F_o,$$

$$[(y/w) - 1 - (F_o/x)]x^2 \to F_1,$$

etc. We have asymptotic convergence.

We could also use

$$\bar{w} = x^{-\frac{1}{2}} \exp(-ix^2/2).$$

Then $W(w,\bar{w}) = 2i$, and one can get solutions of $y'' + x^2 y = 0$ by rewriting the differential equation as

$$y'' + (x^2 - \frac{3}{4x^2})y = -\frac{3}{4x^2} y$$

and applying the formula

$$y(x) = w(x) + (1/2i) \int_x^\infty (3/4\xi^2) [w(x)\bar{w}(\xi) - w(\xi)\bar{w}(x)]\, y(\xi)\, d\xi.$$

We have no guarantee of convergence because we have no control over the size of $y(\xi)$, but

$$F(x) = 1 + \frac{1}{2i} \int_x^\infty \frac{3}{4\xi^2} \left[\bar{w}(\xi)w(\xi) - \frac{w^2(\xi)\bar{w}(x)}{w(x)}\right] F(\xi)\, d\xi,$$

and we can show that F is bounded. This gives an approximation which is good only for large x.

Langer has improved on this method for equations of the type

$$y'' + \varphi^2 y = 0.$$

He has obtained an approximation to the solution expressed in terms of the Bessel function $H_{1/3}$ and φ, which is good for both large and small values of φ.

One can also use the WKB method.

Most differential equations we encounter have irregular singular points at infinity, requiring asymptotic methods to find eigenvalues when the problem involves an infinite interval.

Another example. $y'' + (1/x)\, y = 0$.
Use $w = x^{\frac{1}{4}} \exp(\pm 2i\, x^{\frac{1}{2}})$.

We observe that for the Euler equation $y'' + (1/x^2)\,y = 0$, the asymptotic solution is an exact solution.

DIFFERENTIAL EQUATIONS

Lecture 12: Mass Action

Two substances X and Y react to form a third substance Z. The rate at which Z is formed is given by the law of mass action: The rate of change of the amount of Z, say z, is proportional to the product of the amounts of X and Y as yet untransformed.

Suppose we take x grams of X and y grams of Y and combine them in such a way as to form the compound Z. At first (t = 0) we have 0 grams of Z, but as t increases and the substances X and Y combine to form Z, we have an increasing amount of Z. If α grams of X and β grams of Y yield $\alpha + \beta$ grams of Z, then at any time t, z(t) is made up proportionately of $\frac{\alpha z}{\alpha+\beta}$ grams of X and $\frac{\beta z}{\alpha+\beta}$ grams of Y. So we have $x - \frac{\alpha z}{\alpha+\beta}$ grams of X and $y - \frac{\beta z}{\alpha+\beta}$ grams of Y reacting. Therefore we have the differential equation

$$\frac{dz}{dt} = k\left(x - \frac{\alpha z}{\alpha+\beta}\right)\left(y - \frac{\beta z}{\alpha+\beta}\right) \tag{1}$$

with the initial condition

$$z(0) = 0.$$

If we simplify by letting $K = \frac{kx\beta}{(\alpha+\beta)^2}$, $A = \frac{\alpha+\beta}{\alpha}x$, and $B = \frac{\alpha+\beta}{\beta}y$, we have

$$\frac{dz}{dt} = K(A - z)(B - z). \tag{2}$$

The form of the solution of (2) depends on whether $A = B$ or $A \neq B$ (say without loss of generality that $A > B$).

Suppose $A = B$. Then it is apparent that the solution to (2) is

$$Z = A - \frac{1}{Kt + C} \quad \text{for some } C \in R\text{, where } R = \text{set of real numbers.}$$

The constant C is found to be $\frac{1}{A}$ by applying $z(0) = 0$. Substitution of the original quantities back into the solution gives

$$Z = \frac{\alpha+\beta}{\alpha}x - \frac{\beta kxt}{(\alpha+\beta) + \beta kxt}\,. \tag{3}$$

Notice that $Z \to \frac{\alpha+\beta}{\alpha} x$ as $t \to \infty$.

Now if $A \neq B$ $(A > B)$, we solve (2) and have

$$Z = \frac{BCe^{K(A - B)t} - A}{Ce^{K(A - B)t} - 1} \text{ for some } C \in \mathbb{R}.$$

The condition $z(0) = 0$ implies $C = \frac{A}{B}$, so

$$z = \frac{AB[1 - e^{K(A - B)t}]}{B - Ae^{K(A - B)t}} = \frac{AB[1 - e^{-K(A - B)t}]}{A - Be^{-K(A - B)t}} . \tag{4}$$

Here we note that $z \to B$ as $t \to \infty$, which is the same as in the case $A = B$. This tells us that the termination of the reaction is dependent upon the amount of the quantity Y and the ratio of the reactants.

Suppose we have two substances which combine in a 1:2 ratio, by weight, to form a third substance, and it is observed that 10 minutes after 10 grams of the first substance and 20 grams of the second are mixed, the amount of the product which has been formed is 5 grams. We will find an expression for the amount of the product at any time t.

We have $x = 10$, $y = 20$, $A = (1 + \frac{\beta}{\alpha})\, 10 = 30$, and $B = (1 + \frac{\alpha}{\beta})\, 20 = 30$. So $A = B$. From (3), then, $z = 3 \cdot 10\left(\frac{k \cdot 10 \cdot t}{3/2 + k \cdot 10 \cdot t}\right)$, so $k = 3 \times 10^{-3}$ and $Z = \frac{.6t}{1 + .02t}$ for $t > -50$. Had we mixed 20 grams of each, we would have had

$$A = (1 + \frac{\beta}{\alpha})x = 3 \cdot 20 = 60$$

$$B = (1 + \frac{\alpha}{\beta})y = \frac{3}{2} \cdot 20 = 30.$$

So our solution is given by equation (4):

$$\begin{aligned} z &= \frac{AB[1 - e^{-K(A - B)t}]}{A - Be^{-K(A - B)t}} \\ &= \frac{1800[1 - e^{-\frac{2}{9}k(30)t}]}{60 - 30e^{-\frac{2}{9}k(30)t}} \end{aligned}$$

And since 5 grams of the product are formed after ten minutes,

$$k = \frac{3}{200} \log (1.1).$$

Therefore

$$z = \frac{60\left[1 - e^{-\frac{1}{10}\left(\log \frac{11}{10}\right)t}\right]}{2 - e^{-\frac{1}{10}\left(\log \frac{11}{10}\right)t}} \text{ for all } t.$$

Another way of wording this problem is in terms of dissolving a solid substance in a solution. Consider the following example:

The rate at which a solid substance dissolves varies directly as the amount of undissolved solid present in the solvent and as the difference between the instantaneous concentration and the saturation concentration of the substance. Twenty pounds of solute is dumped into a tank containing 120 lbs. of solvent, and at the end of 12 minutes the concentration is observed to be 1 part in 30. Find the amount of solvent in solution at any time t if the saturation concentration is 1 part of solute to 3 parts of solvent.

If Q is the amount of the material in solution at time t, then $20 - Q$ is the amount of undissolved material at that time and $Q/120$ is the corresponding concentration. Therefore

$$\frac{dQ}{dt} = k(20 - Q)\left(\frac{1}{3} - \frac{Q}{120}\right), \tag{1}$$

with the condition $Q(0) = 0$. Integrating (1) gives

$$-\log (20 - Q) + \log (40 - Q) = \frac{k}{6}t + C.$$

Applying the initial condition $Q(0) = 0$ determines $C = \log 2$. So we write

$$\log \frac{40 - Q}{2(20 - Q)} = \frac{k}{6}t.$$

Applying the results of the observation after 12 minutes determines k to be

$$k = \tfrac{1}{2} \log 9/8.$$

Solving for Q, then,

$$Q = \frac{40\left[1 - e^{-\frac{1}{12}\left(\log \frac{9}{8}\right)t}\right]}{2 - e^{-\frac{1}{12}\left(\log \frac{9}{8}\right)t}} .$$

Note that $Q \to 20$ as $t \to \infty$.

[References: Golomb and Shanks [14], p. 103, and Wylie [32], pp. 24-27.]

Lecture 13: Heat Transfer

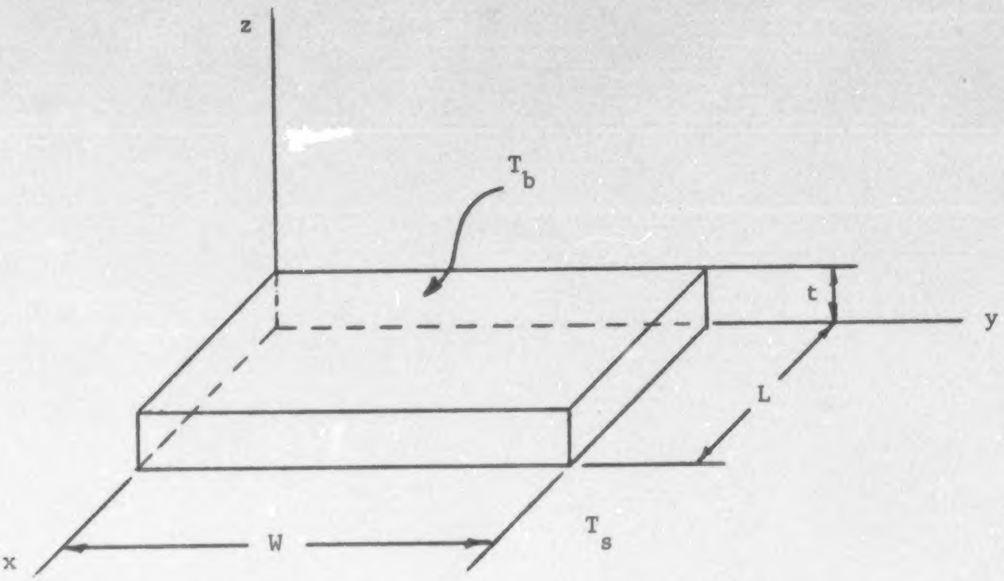
In the absence of a heat-absorbing atmosphere in outer space, many space vehicles are faced with the problem of the dissipation of waste heat. Although in special cases some excess heat may be removed through absorption in the fuel, the prevailing means of dissipation is by radiation. The structure used to radiate the heat is called a radiator fin.

The amount of heat that will be radiated from a surface such as a radiator fin is given by Stefan's Law. The amount of heat dQ radiated from a given surface area dA in a given time interval $d\tau$ is

$$dQ = Es(T^4 - T_s{}^4)\ dA\ d\tau .$$

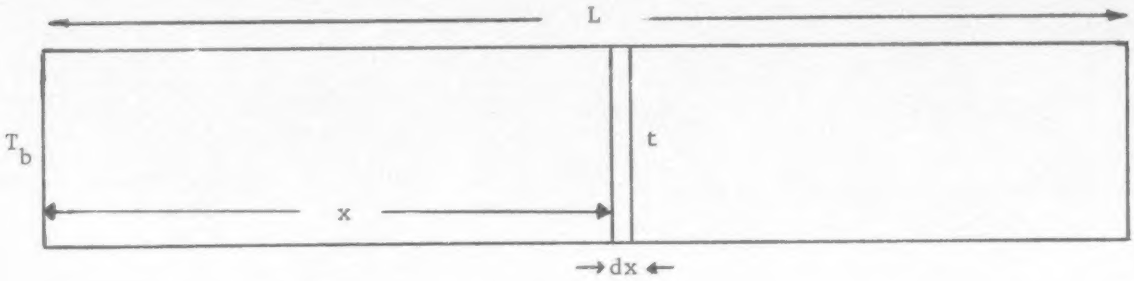
The letter E stands for the emissivity of the surface; s is the Stefan-Boltzmann constant; and T and T_s are respectively the absolute temperatures of the surface and the surrounding medium. The quantity $EsT_s{}^4\ dA\ d\tau$ is the incident thermal radiant energy that is absorbed from the sun and other objects from which the fin will receive radiant heat. In order to find the amount of heat that a fin will radiate, we first need to know the temperature T at any point (x,y,z) of the fin.

We consider a metal plate of length L, width W, thickness t, mean conductivity k, emittance E, supplied by an infinite heat source at temperature T_b, losing heat by radiation only (operating in a vacuum) with sink temperature T_s. See figure on next page.



If we make the assumption that the thickness t of a fin is small, so that the temperature is essentially constant in the z-direction, and that the heat lost from the edges y = 0, y = W, and x = L is negligible, then we have a one-dimensional heat flow problem. This is because (1) the temperature is constant in the z-direction and (2) the symmetry of the boundary conditions for y = 0 and y = W implies that the heat does not flow in the y-direction.

Consider the volume element tw dx in the figure below:



The heat conducted into the element at the point x, denoted by $Q_c(x)$, is

$$Q_c(x) = -ktw \frac{dT}{dx}.$$

At x + dx, the heat conducted is

$$Q_c(x + dx) = -ktw \frac{dT}{dx} + \frac{d}{dx}\left(Q_c(x)\right)dx.$$

The heat radiated from the surface between x and x + dx is

$$Q_r(x) = snE(T^4 - T_s^4)w\,dx$$

(n is the number of sides radiating: n = 1 or 2). Since the energy of the system must be conserved,

$$Q_c(x) = Q_c(x + dx) + Q_r(x)$$

or

$$\frac{d}{dx}\left[-ktw\,\frac{dT}{dx}\right] dx + snE(T^4 - T_s^4)w\,dx = 0.$$

Simplifying, we obtain

$$\frac{d^2T}{dx^2} = \frac{nsE}{kt}\,(T^4 - T_s^4). \tag{1}$$

This is the equation for all x such that $0 < x < L$. At x = 0,

$$T(0) = T_b.$$

Because we are assuming that the heat lost at x = L is negligible,

$$\frac{dT}{dx} = 0 \text{ at } x = L.$$

Equation (1) can be integrated at once if we multiply by $2\,\frac{dT}{dx}$, yielding

$$\left(\frac{dT}{dx}\right)^2 = \frac{2snE}{5kt}\,(T^5 - 5T_s^4T) + C. \tag{2}$$

Evaluating this at x = L gives

$$C = -\frac{2snE}{5kt}\left[T^5(L) - 5T_s^4T(L)\right].$$

Since T(x) is a decreasing function, $\frac{dT}{dx} < 0$. Therefore

$$\frac{dT}{dx} = -\sqrt{\frac{2nsE}{5kt}\,(T^5 - 5T_s^4T) + C}\ . \tag{3}$$

Separating variables in equation (3) and integrating with x ranging from 0 to L and T from T_b to T(L):

$$\int_{T_b}^{T(L)} \frac{dT}{\sqrt{\frac{2nsE}{5kt}\,(T^5 - 5T_s^4T - T^5(L) + 5T_s^4(L)}} = -L$$

Equation (4) is solved for the parameter T(L) by numerical methods. Knowing T(L), the solution to (1) is

$$\int_{T_b}^{T} \frac{dT}{\sqrt{T^5 - 5T_s^4 T - T^5(L) + 5T_s^4 T(L)}} = -x\sqrt{\frac{2nsE}{5kt}} .$$

If we make the additional assumptions that $T_s = 0$ and the length L of the fin is infinite, in which case $T(L) = T_s = 0$, equation (3) may be written as

$$\frac{dT}{dx} = -\sqrt{\frac{2nsE}{5kt}}\, T^{5/2} .$$

Once again we integrate and obtain

$$-\frac{2}{3}\left[T^{-3/2} - T_b^{-3/2}\right] = -x\frac{2nsE}{5kt} .$$

Solving for T, we have

$$T = \left[\frac{1}{\frac{3}{2}\, x\sqrt{\frac{2nsE}{5kt}} + T_b^{-3/2}}\right]^{2/3} .$$

References: Bartas and Sellers [4], Carslaw and Jaeger [8], Liebleim [21], and Sears and Zemansky [29].

DIFFERENTIAL EQUATIONS

References

1. Agnew, Ralph P. *Differential Equations*, 2nd ed. New York, McGraw-Hill Book Company, 1962.

2. Allen, Roy G. *Mathematical Economics*, 2nd ed. New York, St. Martin's Press, Inc., 1959.

3. Bailey, N. T. *The Mathematical Approach to Biology and Medicine*. New York, John Wiley and Sons, Inc., 1967.

4. Bartas, J. G. and Sellers, W. S. "Radiation Fin Effectiveness." *Journal of Heat Transfer*, Trans. ASME, Series C, 82 (1960), pp. 73-75.

5. Baumol, William J. *Economic Dynamics*, 3rd ed. New York, The Macmillan Company, 1970.

6. Bishir, John W. and Drewes, Donald W. *Mathematics in the Behavioral and Social Sciences*. New York, Harcourt Brace Jovanovitch, Inc., 1970.

7. Bôcher, M. *Leçons sur les méthodes de Sturm*. Paris, Gauthier-Villars, 1917.

8. Carslaw, Horatio S. and Jaeger, J. C. *Conduction of Heat in Solids*, 2nd ed. New York, Oxford University Press, 1959.

9. Coleman, James S. *Introduction to Mathematical Sociology*. New York, Free Press (Division of the Macmillan Company), 1964.

10. Dorf, Richard C. *Modern Control Systems*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1967.

11. Feynman, Richard P., *et al*. *The Feynman Lectures on Physics*, Vol. I. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1963.

12. Gantmacher, Felix R. *Theory of Matrices*, Vol. II. New York, Chelsea Publishing Company, 1959.

13. Goldstein, Herbert. *Classical Mechanics*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1950.

14. Golomb, Michael and Shanks, M. E. *Elements of Ordinary Differential Equations*, 2nd ed. New York, McGraw-Hill Book Company, 1965.

15. Kaplan, Wilfred. *Advanced Calculus*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1952.

16. Kaplan, Wilfred. *Operational Methods for Linear Systems*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1962.

17. Kaplan, Wilfred. *Ordinary Differential Equations*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1958.

18. Kooros, A. *Elements of Mathematical Economics*. Boston, Massachusetts, Houghton Mifflin Company, 1965.

19. Langer, Rudolph E. *Ordinary Differential Equations*. New York, John Wiley and Sons, Inc., 1954.

20. Licklider, J. C. R. "Quasi-linear Operator Models in the study of Manual Tracking." *Developments in Mathematical Sociology*. Glencoe, Illinois, Free Press, 1960, pp. 171-280.

21. Liebleim, S. "Analysis of a Temperature Distribution and Radiant Heat Transfer along a Rectangular Fin of Constant Thickness." NASA TN D-196, November, 1959.

22. Martin, William T. and Reissner, E. *Elementary Differential Equations*, 2nd ed. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1961.

23. Moulton, Forest R. *Introduction to Celestial Mechanics*. New York, Dover Publications, 1970.

24. Perlis, Sam. *Theory of Matrices*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1952.

25. Phillips, Henry B. *Differential Equations*, 3rd ed. New York, John Wiley and Sons, Inc., 1946.

26. Pontrjagin, L. S. *Ordinary Differential Equations*. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1962.

27. Ritger, Paul D. and Rose, Nicholas J. *Differential Equations with Applications*. New York, McGraw-Hill Book Company, 1968.

28. Scarborough, James B. *Differential Equations and Applications*. Baltimore, Maryland, Waverly Press, Inc., 1965.

29. Sears, F. W. and Zemansky, M. W. *University Physics*, 3rd ed. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., 1964.

30. Thrall, Robert M., *et al*., editors. *Some Mathematical Models in Biology*, rev. ed. [Available from National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia 22151.]

31. von Kármán, T. and Biot, M. *Mathematical Methods in Engineering*. New York, McGraw-Hill Book Company, 1940.

32. Wylie, Clarence R. Advanced Engineering Mathematics, 3rd ed. New York, McGraw-Hill Book Company, 1966.

33. Mathematics of the Decision Sciences, part 2. Lectures in Applied Mathematics, vol. 12. Providence, Rhode Island, American Mathematical Society, 1968.

34. Homeostasis and Feedback Mechanisms. Proceedings of the 18th Symposium of the Society for Experimental Biology. New York, Academic Press, Inc., 1965.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 1: Binomial Probabilities and Applications

A. Football Parlay

It seems that probability may have been of interest first to the gambling operators and, no doubt, is yet of interest to gamblers, but today I'd like to turn the tables to see whether an elementary discussion of a big gambling enterprise will help to reduce the "take." Consider the handout which would tempt you to pick the winners and win a big pot. At the cost of one dollar you attempt to win by picking as indicated, according to the legend on the handout [page 287], e.g., pick 4 winners in 4 tries to be paid $9.00 or pick 5 winners in 5 tries to be paid $15.00, etc. Answer the following questions.

1. What is the probability of winning if you pick 4 games? What is your expected net return? (Repeat using $p = 0.60$, and determine p to break even.)
2. What is the probability of making 10 correct choices in 10 tries? 9 correct in 10 tries? What is the probability of winning with 10 tries?
3. Suppose that, instead of making 10 tries, you just decide to donate your expected loss to the group running the show. How much should you donate?

To answer these questions we need to characterize the problem and perhaps introduce some assumptions. We note:

a) There are only two possible outcomes for each pick--win or lose.

b) Trials are independent. That is to say, picking a winner in one game does not influence other selections.

c) What are the odds on picking a winner? The point spreads are supposedly such that each game is a toss-up, but ties count as losses. Let us assume the possibility of a correct pick is 0.50--constant for each selection.

These conditions with the fixed number (10) of trials meet all the requirements of the binomial probability model. We may now attack the proposed questions using the binomial probability model, which is given by

$$f(x) = \binom{n}{x} p^{x}(1 - p)^{n-x}, \quad x = 0, 1, 2, \ldots, n,$$

with x = number of successes in n trials, p = probability of a success on each trial.

SATURDAY, OCTOBER 10, 1970

4 Teams	9	Ties Lose	8 Teams	61
5 Teams	15		9 Teams	101
6 Teams	25		10 Teams	171
7 Teams	41		9 out of 10	15

1	Auburn		27	Wisconsin	
2	Clemson	Plus 17	28	Iowa	Plus 3
3	Penn State		29	Kansas State	
4	Boston College	Plus 10	30	Kansas	Plus 7
5	West Virginia		31	Nebraska	
6	Duke	Plus 14	32	Missouri	Plus 7
7	Michigan		33	Oklahoma State	
8	Purdue	Plus 3	34	TCU	Plus 7
9	No. Carolina		35	Air Force	
10	So. Carolina	Plus 3	36	Tulane	Plus 14
11	Texas		37	Colorado	
12	Oklahoma	Plus 17	38	Iowa State	Plus 14
13	Wake Forest		39	Washington	
14	Virginia Tech	Plus 7	40	California	Plus 3
15	Florida		41	So. California	
16	Florida State	Plus 6	42	Stanford	Plus 3
17	Mississippi		43	Texas Tech	Even
18	Georgia	Plus 7	44	Texas A.M.	
19	Tennessee		45	UCLA	
20	Georgia Tech	Plus 3	46	Oregon	Plus 10
21	Houston		47		
22	Miss. State	Plus 10	48		
23	Northwestern		49		
24	Illinois	Plus 3	50		
25	Minnesota		51		
26	Indiana	Plus 7	52		

B. Application to Comparative Experiments

Twenty strips will be painted across a road with paint A and twenty will be painted with paint B. Then, after a period of exposure to traffic and weather, the Highway Supervisor will select the better paint. Some problems:

1. It would be most convenient to paint first 20 strips of paint A and then 20 of paint B. Comment on the procedure.

2. Suppose we'd rather choose 20 blocks of 2 strips and randomly assign A or B to the strips. Then the supervisor says for each pair, "This is the better." If in fact the paints are equally good, what is the probability that paint A is declared to be the better in 15 or more cases? What is the probability that one or the other will be declared better in 15 or more cases? If A does in fact win 15 out of 20, at what level of significance is A the better?

3. If in fact the probability is .6 that on a given pair paint A will be better than paint B, what is the probability that paint B would appear better in our experiment?

One can see a number of drawbacks to the experimental layout suggested in problem 1. For example, there may be an oil slick on the road surface where paint A is applied but not where B is used. If this is the case, the two brands would be observed under different road conditions. To overcome this type of difficulty, one could apply the two brands of paint alternately, but does this procedure solve the problem? To answer this, we must consider the population of inference: all roads to be painted.

A solution to the experimental layout is given in problem 2 above. We see that the two brands are to be compared in blocks to ensure that each pair is observed under identical conditions. Also, the twenty blocks or test sites are chosen to represent the entire population. Fair judging is assured by not identifying brands within blocks to the supervisor.

The experiment may now be characterized as consisting of twenty independent trials, each having two possible outcomes with a constant probability of a brand's being declared the better. Again, the requirements of the binomial probability model are fulfilled, and problems 2 and 3 may now be solved.

C. Other Settings for Binomial Problems

1. A liquid fertilizer is manufactured locally. The manufacturer donated a quantity of his fertilizer to the Forestry Service and later asked for an evaluation of its effectiveness. No evaluation could be made, as a controlled experiment was not conducted.

An evaluation based on sound probabilistic considerations could have been made. This problem is quite similar to part B above. Pairs of equivalent trees should have been selected. Randomly administer equal amounts of liquid fertilizer and water. Determine qualitatively which of the pair has performed the better. Careful planning again allows us to use the binomial probability model.

2. This year I planted two varieties of strawberries as two separate plots. Unfortunately, I will not be able to judge which of the two varieties is the better. The experimental layout is poor, as the two varieties are compared under different conditions, i.e., two different plots are involved.

To overcome this difficulty, I could have arranged an experimental layout in pairs. However, there is another problem. How does one compare two varieties of strawberries? The two will ripen at different times, last for different periods, and differ somewhat in size and flavor. Perhaps a reasonable measure for comparison is the volume of berries produced. We now are considering a quantitative measure--more information is available than is needed for the qualitative binomial model. Perhaps this added information could be used by introducing some other model.

3. Watermelons are planted in a hill. The seed package states that a seed will germinate with probability 0.80. May the binomial probability model be used to evaluate the probability that at least three out of 10 seeds planted in a hill will germinate? Answer: No; the seeds are planted together, so the independence requirement is violated.

4. Three years from next spring I hope to have 100 apple trees in a row. Knowing the probability that a tree survives three years is 0.70 and letting x be the number of living trees at the end of the three years, determine the number of trees which I should order so that

$$Pr(x \geqq 100) = .95.$$

Answers:

Section A

1. With $p = .5$

a) 1/16

b) $\frac{1}{16}(8) + \frac{15}{16}(-1) = -.4375$ (a loss)

With $p = .6$

a) $.6^4 = .1296$

b) $.1296(8) + (.8704)(-1) = .1664$ (a gain)

To break even

$$0 = p^4(8) + (1 - p^4)(-1)$$

or

$$p = 1/\sqrt{3} \cong .577$$

2. a) $\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}$

b) $10\left(\frac{1}{2}\right)^{10} = \frac{10}{1024}$

c) $\frac{11}{1024}$

3. The expected net return on a dollar bet is

$$\frac{1}{1024}(170) + \frac{10}{1024}(14) + \frac{1013}{1024}(-1) = -.686.$$

Therefore you should donate 69 cents.

Section B

2. a) $\sum_{x=15}^{20} \binom{20}{x}\left(\frac{1}{2}\right)^{20} = .0207$

b) 2(.0207)

c) L.S. = the probability that one will be at least this much better than the other, given that they are equally good = 2(.0207)

3. $P(11 \text{ or more favor } B \mid p = .4) = .1275$

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 2: The Poisson Probability Law and Applications

Problem. A newsboy sells papers at the corner every Sunday. He knows from past experience that he could sell, on the average, 20 copies of the New York Times at 75¢ each if he has an unlimited supply (i.e., the average demand for the New York Times is 20). The papers cost him 60¢ each. How many should he order?

It is often reasonable to assume that the random variable called the demand (say x) follows a Poisson model with parameter λ equal to the average demand. The Poisson model is given by

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

and is generally useful for situations where (1) there are many opportunities for the event to occur and (2) for each opportunity the probability of occurrence is small.

Suppose the newsboy orders N papers, and let $\mathcal{P}(x)$ be the profit if x papers are demanded. Then

$$\mathcal{P}(x) = \begin{cases} .15x - .60(N - x) & \text{if } x \leq N \\ .15N & \text{if } x > N \end{cases}$$

and

$$E_N[\mathcal{P}(x)] = \sum_{x=0}^{N} (.75x - .60N)\,\frac{e^{-20}(20)^{x}}{x!} + \sum_{x=N+1}^{\infty} .15N\,\frac{e^{-20}(20)^{x}}{x!}\,.$$

Now, we could evaluate $E_N[\mathcal{P}(x)]$ for several values of N and thereby choose the value of N that maximizes the expected profit. However, let us try the following approach. Suppose the boy looks at the expected profit from ordering the $(N + 1)^{st}$ paper. This profit from ordering the $(N + 1)^{st}$ paper will be

$$\begin{aligned} &.15 \text{ if } x > N \\ -&.60 \text{ if } x \leq N, \end{aligned}$$

and hence

$$\begin{aligned} E[\text{profit from } (N + 1)^{st} \text{ paper}] &= -.60\, P(x \leq N) + .15\, P(x > N) \\ &= -.6 + .75\, P(x > N). \end{aligned}$$

We easily find

$$E[\text{profit from } (N + 1)^{st} \text{ paper}] = \begin{cases} .0327 & \text{if } N = 16 \\ -.01525 & \text{if } N = 17, \end{cases}$$

so the newsboy should order 17 papers.

This same kind of analysis would apply to all problems in which there is a demand random variable which it is reasonable to assume is a Poisson random variable. How many hamburgers shall we prepare for the noon rush? How many quarts of milk shall we stock in a machine if we discard the old daily? How much sweet corn should the kids pull to stock their roadside stand? How many tickets should a scalper buy?

We might ask a different type of question even when the same model is appropriate. We may ask, "How many parts of a certain kind should a supply house stock in order to meet the month's demand with probability of .90? Assuming the sales records from past years indicate an average monthly demand of 5 parts, how many should we have on hand at the beginning of the month?" A check of the Poisson tables gives

$$P(9 \text{ or more calls}) = .0680$$

$$P(8 \text{ or more calls}) = .1333$$

Thus, we would stock 8 such parts.

Suppose we are building a new hospital in State College to serve 40,000 people whose birth rate is .02 (per year per person). That is, there is an average of 800 babies born per year. How many baby beds shall we provide in the nursery? Let us assume that each baby stays 4 days and that we want to determine a number N of beds such that for any given time, if there are x babies in the hospital, we have

$$P(x \leq N) = .99.$$

Solution: Using the Poisson model with an average demand

$$\lambda = \frac{40,000\ (.02)\ 4}{365} \cong 9, \text{ we get } N = 17.$$

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 3: Bayes' Theorem and Applications

A. The Oil Exploration Problem. Suppose it is true that if you drill at a random location in a certain Oklahoma county, the probability that you strike oil is .1. However, it has been observed that if there is oil, the probability of a certain rock formation's being present is .5. It is also observed that the probability of the occurrence of this rock formation when there is nothing of value beneath is .2. What is the probability of oil beneath a spot with this rock formation?

Solution:

$$P(O|R) = \frac{P(O \cap R)}{P(R)}$$

$$= \frac{P(O)\ P(R|O)}{P(O)\ P(R|O) + P(\overline{O})\ P(R|\overline{O})}$$

$$= \frac{(.1)\ (.5)}{(.1)\ (.5) + (.9)\ (.2)} = \frac{5}{23} \cong .217.$$

This is an application of Bayes' Theorem:

Let $A_1, A_2, \ldots, A_k$ be a mutually exclusive and exhaustive set of outcomes of a random process, and let B be a chance event such that $P(B) \neq 0$. Then

$$P(A_i|B) = \frac{P(B|A_i)\ P(A_i)}{\sum_{j=1}^{k} P(B|A_j)\ P(A_j)}.$$

B. Suppose a test for diagnosing tuberculosis has the properties that a patient with TB will have a positive response with probability of .95 and a patient without TB will have a positive response with probability of .02. If the incidence of TB is .001 and a person selected at random has a positive response (PR), what is the probability he has TB?

Solution:

$$P(TB|PR) = \frac{P(TB)\ P(PR|TB)}{P(TB)\ P(PR|TB) + P(\overline{TB})\ P(PR|\overline{TB})}$$

$$= \frac{(.001)\ (.95)}{(.001)\ (.95) + (.999)\ (.02)}$$

$$= \frac{95}{2093} \cong .045.$$

This result may seem low, however, since the test seems rather good. The explanation is that when a positive response is found, it is much more likely due to a machine malfunction than to the fact that the person is tubercular.

C. The Car Trouble Problem. The manager of a car repair shop knows from past experience that when called by a woman whose car will not start, the probabilities for various troubles are as follows:

	Trouble	Probability
F	Flooded	.3
BC	Battery cable loose	.2
PB	Points bad	.1
OG	Out of gas	.3
SE	Something else	.1

He also knows that if the woman will hold the gas pedal down and try to start the car for 30 seconds, the probability that it will start (S) is

$P(S|F) = .9$ $P(S|BC) = 0$ $P(S|PB) = .2$

$P(S|OG) = 0$ $P(S|SE) = .2$

a) Given that a lady has called and is instructed to "hold the pedal...", what is the probability that her car will start?

b) Given that her car does start when she "holds the pedal...", what is the probability that her car was flooded?

D. Bayesian Diagnosis of Diseases. The problem of diagnosing a patient's illness is being considered by the statistician. We'll work our way up toward the real problem by considering some trivial cases.

(a) The country with just one disease. If you lived in a country with just one disease, then it would be true that you could make no mistake in diagnosing the patient's illness, assuming (and this might be a big assumption) that he really is ill. We would say that the prior probability of this disease, say X, is 1 for a patient who really is ill.

(b) The country with two diseases. If you lived in a country with two diseases, say X and Y, and if in the past 90 per cent of those ill had disease X and 10 per cent had disease Y, then we would assign prior probabilities for these diseases as P(X) = .9 and P(Y) = .1. Assuming no changes in these figures, we would be correct .9 of the time if we simply told each patient he had illness X. However, since we are still supposing, let us also suppose that there are observations A, B, C, ... that we can make on the patient, and we'll give low, medium, high as the possible observation for each of these. Suppose it is known that disease X always results in a high observation of A and disease Y always results in a low reading for A. Then again you can diagnose without error.

Let us make the game more interesting by saying that the prior for X is .9 and for Y is .1, but the probabilities for the A observation are as in the table below.

Patient with	Observation A		
	Low	Medium	High
X	.1	.3	.6
Y	.6	.3	.1

Conditional Probabilities for Diseases X and Y, given A

We would now revise the probabilities for diseases X and Y after observing A:

Observed A	Conditional Probability X	Conditional Probability Y	Diagnosis
Low	9/15	6/15	X??
Medium	9/10	1/10	X?
High	54/55	1/55	X
Prior Probabilities	.9	.1	

The strong effect of the prior probabilities for diseases X and Y is apparent in the above table. Now, suppose that diseases X and Y occur equally often. Then, with the probabilities for the levels of A as before, we would have

Observed A	Conditional Probability X	Conditional Probability Y	Diagnosis
Low	1/7	6/7	Y?
Medium	1/2	1/2	??
High	6/7	1/7	X?
Prior Probabilities	.5	.5	

The formula used in calculating these new conditional probabilities is simply

$$P(X|A) = \frac{P(X)\ P(A|X)}{P(X)\ P(A|X) + P(Y)\ P(A|Y)} .$$

(c) Two diseases, two observations, equal priors. We shall be able to improve our probability for correct diagnosis if we make more observations on the patient. This time let us suppose there are diseases X and Y, again with prior probabilities of .5 for each, and that the probabilities for low, medium, and high for the observations A and B are as given in the table below.

Patient with	Observation A			Observation B		
	Low	Medium	High	Low	Medium	High
X	.1	.3	.6	.2	.2	.6
Y	.6	.3	.1	.6	.3	.1

Using this table one would no doubt classify a patient with A low and B low as having disease Y. The following data were generated artifically according to the probabilities given above. Classify each as belonging to disease X or Y.

Patient	A	B	Disease
1	High	Low	
2	Low	Low	
3	High	Medium	
4	Low	Medium	
5	High	Medium	
6	Low	Low	
7	Low	Low	
8	Medium	High	
9	Low	Low	
10	Low	Low	
11	Medium	Low	
12	Low	Medium	
13	High	Low	
14	High	High	

Patient	A	B	Disease
15	Medium	High	
16	Low	Low	
17	High	High	
18	Low	Medium	
19	High	Low	
20	Medium	High	

The statistician actually calculates conditional probabilities for diseases X and Y given the observations A and B. The table below displays these probabilities for the problem which you have just worked. Do you suppose the doctor should give his patient a report in terms of probabilities, just as the weatherman now does?

Observed		Conditional Probability		Diagnosis
A	B	X	Y	
Low	Low	2/38	36/38	Y
Low	Medium	1/10	9/10	Y?
Low	High	1/2	1/2	??
Medium	Low	1/4	3/4	Y?
Medium	Medium	2/5	3/5	??
Medium	High	6/7	1/7	X?
High	Low	2/3	1/3	X?
High	Medium	4/5	1/5	X?
High	High	36/37	1/37	X

The statistician would have classified these 20 patients according to the most probable disease, flipping a coin when the conditional probability is .5 for each disease. The results for these 20 patients and the actual disease assumed for each patient is given on the next page.

Patient	A	B	Statistician	Actual
1	High	Low	X	X
2	Low	Low	Y	Y
3	High	Medium	X	X
4	Low	Medium	Y	Y
5	High	Medium	X	X
6	Low	Low	Y	Y
7	Low	Low	Y	Y
8	Medium	High	X	Y
9	Low	Low	Y	Y
10	Low	Low	Y	Y
11	Medium	Low	Y	Y
12	Low	Medium	Y	Y
13	High	Low	X	X
14	High	High	X	X
15	Medium	High	X	X
16	Low	Low	Y	Y
17	High	High	X	X
18	Low	Medium	Y	Y
19	High	Low	X	X
20	Medium	High	X	X

Thus the statistician is correct in 19 of 20 times. The probability that the statistician is wrong in his diagnosis of an individual can be calculated and is found to be .15. Thus the statistician was quite lucky this time; he will, on the average, have 3 incorrect diagnoses in 20 for the situation described in this section.

(d) Two diseases, two observations, unequal priors. For our last game, suppose disease X has prior probability of .2 and disease Y has prior probability of .8. Then in 20 patients we would expect to find 4 with disease X and 16 with disease Y. We wouldn't do too badly if we said that everyone had disease Y. Try to classify the following 20 patients.

Patient	A	B	Disease
1	Low	Low	
2	Low	Medium	
3	Low	Low	
4	Medium	High	
5	Low	Low	
6	High	Medium	
7	Low	Low	
8	Low	Low	
9	Low	Medium	
10	Low	Medium	
11	Medium	High	
12	High	High	
13	High	Low	
14	Low	High	
15	Low	Low	
16	Low	Low	
17	Low	Medium	
18	Medium	Low	
19	Low	Low	
20	High	High	

With the priors of .2 for X and .8 for Y, the conditional probabilities for diseases X and Y, given observations A and B, are as given on the next page.

Observed		Conditional Probability		Diagnosis
A	B	X	Y	
Low	Low	1/73	72/73	Y
Low	Medium	1/37	36/37	Y
Low	High	1/5	4/5	Y?
Medium	Low	1/13	12/13	Y
Medium	Medium	1/7	6/7	Y?
Medium	High	3/5	2/5	X?
High	Low	1/3	2/3	Y?
High	Medium	1/2	1/2	??
High	High	9/10	1/10	X

In this case the statistician would classify patients 4, 11, 12, and 20 as having disease X and with probability of .5 he would declare patient 6 to have disease X. Actually, patients 6, 11, 12, and 20 were assigned disease X and all others were assigned Y, so this time the statistician has 1.5 errors out of 20, i.e., an error rate of 7.5 per cent. Actually, he was lucky again, as his real error rate for this problem should be 12.4 per cent. It will be noted that the error rate has decreased due to a more discriminating (or more informative) prior for diseases X and Y.

(e) Questions:

1. How do we obtain the probabilities called for by the above Bayesian analysis?

2. Among the possible tests that one may call for a given patient, how do we select the one to run next?

3. Is this type of analysis actually the process which the doctor uses? Could or should he be a bit more objective in his calculations?

4. Would it be helpful to read into a computer the priors and the observations and have the calculated conditional probabilities read out?

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 4: Models and Estimation

Let us consider the following situation which might be encountered by an engineer:

During the manufacture of an item, a cylindrical rod with circular cross-section has to be fitted into a circular socket. The rod diameter X has a normal distribution with known mean and variance μ_X and σ_X^2, respectively. The socket diameter Y is also normal, with parameters μ_Y and σ_Y^2. A typical problem associated with this situation is to determine the probability that the rod fits into the socket. Or perhaps one might want to determine, say, μ_X so that the gap or interference meets some engineering specifications. These problems may be solved. Notice, however, that in the statement of the problem we have assumed the normal distribution and also values of the parameters.

Often one would prefer a more flexible model than the symmetric normal. Three models which are often employed when skewness is present are:

1. Gamma

$$f(X;\alpha,\beta) = \frac{X^{\alpha-1} e^{-X/\beta}}{\Gamma(\alpha)\, \beta^{\alpha}}, \qquad X \geqq 0$$

where

$$\Gamma(\alpha) = \int_0^{\infty} X^{\alpha-1} e^{-X}\, dX$$

$$= (\alpha - 1)\, \Gamma(\alpha - 1)$$

2. Weibull

$$f(X;b,c) = \frac{c}{b} \left(\frac{X}{b}\right)^{c-1} e^{-\left(\frac{X}{b}\right)^{c}}, \qquad X \geqq 0$$

which has the nice distribution function

$$F(X) = 1 - e^{-\left(\frac{X}{b}\right)^{c}}.$$

3. Log-Normal

If $X \sim N(\mu,\sigma^2)$, then $Y = e^X \sim$ log-normal (μ,σ^2)

To use these or other models we need to know the parameters. Let's look briefly at some estimation problems. Consider a random sample of $X_1, \ldots, X_n$ from a normal population having unknown parameters μ and σ^2. Three commonly used estimators of σ^2 are

$$\frac{1}{n-1} \sum (X_i - \overline{X})^2,$$

$$\frac{1}{n} \sum (X_i - \overline{X})^2,$$

and

$$\frac{1}{n+1} \sum (X_i - \overline{X})^2.$$

Each of these is prompted by different considerations. To investigate these we need to recall that with the sample as described we have

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1} ,$$

so that

$$E\left(\sum (X_i - \overline{X})^2\right) = (n - 1)\, \sigma^2$$

and

$$\mathrm{var}\left(\sum (X_i - \overline{X})^2\right) = 2(n - 1)\, \sigma^4.$$

With these results we see that the first estimator is unbiased. That is,

$$E\left(\frac{1}{n-1} \sum (X_i - \overline{X})^2\right) = \sigma^2.$$

The second estimator is the maximum likelihood estimator, which we will discuss later. The third estimator is the solution to the following problem.

Suppose $X_1, \ldots, X_n$ is a random sample from a normal population as above. Find A such that the mean square error of $A \sum (X_i - \overline{X})^2$ is a minimum for estimators of σ^2. Recall that if $\hat{\theta}$ is an estimator of θ, then the mean square error (M.S.E.) is defined as

$$\text{M.S.E.}\ (\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

$$= \text{var}\ \hat{\theta} + (\text{Bias}\ \hat{\theta})^2,$$

where Bias $\hat{\theta} = E(\hat{\theta}) - \theta$.

For our problem we then have

$$\text{M.S.E.}\ \left(A \sum (X_i - \overline{X})^2\right) = A^2\ 2(n-1)\sigma^4 + [A(n-1)\sigma^2 - \sigma^2]^2.$$

Minimizing with respect to A, we obtain the desired result.

Two of the most common approaches to estimation problems are the method of moments and maximum likelihood. Let's discuss these techniques assuming we have a random sample from a Weibull population and need to estimate the parameters b and c.

The method of moments involves equating population moments to sample moments and solving. That is, set

$$E(X) = \sum X_i/n$$

$$E(X^2) = \sum X_i^2/n$$

and solve for b and c. We need then the first two moments of X. In the case in which X has a Weibull distribution,

$$E(X^k) = \int_0^\infty \frac{c}{b}\, X^k \left(\frac{X}{b}\right)^{c-1} e^{-\left(\frac{X}{b}\right)^c}\, dX$$

$$\text{Let } Z = \left(\frac{X}{b}\right)^c$$

$$= b^k \int_0^\infty Z^{k/c}\, e^{-Z}\, dZ$$

$$= b^k\, \Gamma\left(1 + \frac{k}{c}\right).$$

We have then

$$E(X) = b\ \Gamma\left(1 + \frac{1}{c}\right)$$

$$E(X^2) = b^2 \Gamma\left(1 + \frac{2}{c}\right)$$

and

$$\operatorname{var} X = E(X^2) - [E(X)]^2$$

$$= b^2 \left[\Gamma\left(1 + \frac{2}{c}\right) - \left\{\Gamma\left(1 + \frac{1}{c}\right)\right\}^2\right].$$

The method of moments requires us to solve the set of equations

$$b\,\Gamma\left(1 + \frac{1}{c}\right) = \sum X_i/n$$

and

$$b^2\,\Gamma\left(1 + \frac{2}{c}\right) = \sum X_i^2/n\ .$$

Eliminating b, we must now solve by numerical techniques

$$\frac{\left[\Gamma\left(1 + \frac{1}{c}\right)\right]^2}{\Gamma\left(1 + \frac{2}{c}\right)} = \frac{\Sigma X_i}{\Sigma X_i^2}\ .$$

Perhaps the most popular method of estimation is the method of maximum likelihood. We define the likelihood function as

$$L(b,c;X_1,\ldots,X_n) = \prod_{i=1}^{n} f(X_i;b,c).$$

The likelihood function is then the joint density of the sample; however, $X_1, \ldots, X_n$ are considered fixed, with b and c being the variables. The maximum likelihood estimators are then the point in the parameter space which maximizes the likelihood function. Here we need to maximize L with respect to b and c. For the Weibull model using maximum likelihood also requires solving a set of nonlinear equations; however, the estimators have many desirable properties. See, for example, the following references:

Freund, John E. Mathematical Statistics. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1971.

Hogg, Robert V. and Craig, Allen T. Introduction to Mathematical Statistics. New York, The Macmillan Company, 1970.

Mood, Alexander M. and Graybill, F. A. Introduction to the Theory of Statistics, 2nd ed. New York, McGraw-Hill Book Company, 1963.

Problems:

1. Find the maximum likelihood estimators for the mean and variance of a normal population.

2. Find the constant K so that $E[K \sum (X_i - \overline{X})^2]^{\frac{1}{2}} = \sigma$, where the X_i are a random sample from a normal population. Hint: Let $u = \sum (X_i - \overline{X})^2/\sigma^2 \sim \chi^2_{n-1}$. Then we need $E(u^{\frac{1}{2}})$, where

$$f(u) = \frac{u^{\frac{1}{2}(n-3)} e^{-u/2}}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{1}{2}(n-1)}}, \quad u \geqq 0.$$

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 5: Confidence Intervals

Let X be a random variable (possibly a vector of random variables) and let L and U be two functions of X. Let θ be some quantity of interest, and suppose

$$P[L(X) \leqq \theta \leqq U(X)] = .95.$$

Then after observing X, say x, the interval (L(x),U(x)) is a .95 confidence interval for θ. This is a mathematical definition of a confidence interval. To say that (L(x),U(x)) is a .95 confidence interval for θ means that intervals obtained as this one was obtained will contain θ 95% of the time.

Now in practice it is often somewhat erroneously said that the "probability that the interval contains θ is .95." We shall see later that the Bayesians can do this and have probability intervals where we must have confidence intervals.

I would like to impress upon you the importance of confidence intervals in statistics. I was once a physics student and always troubled by the way in which experimental results were reported, e.g., 104.27 ± .05. I never could find out what the ± .05 meant, and I couldn't get much help from my instructors. Some would say it was an indication of the error in the measuring devices, others the "maximum possible error," etc. So when introduced to confidence intervals with their precise definition, I was delighted.

The best illustration of the meaning of confidence intervals was given by a professor from Iowa State who was employed by a company to estimate the orange crop in Florida. The first year he gave the company a .67 confidence interval on the orange crop; when the crop was in, the interval contained the value. The next year he again gave a .67 confidence interval on the orange crop and the company believed him and acted accordingly with some constraint. When the crop was in, it was seen that it was within the confidence interval. For the third year the professor again gave a 2/3 confidence interval on the orange crop. The company acted without restraint, but, when the crop was in, it was not contained in the confidence interval--and that was the end of the trips to Florida. However, these intervals had acted exactly as 2/3 confidence intervals will (in the long run); i.e., they contained the true value of the quantity of interest 2/3 of the time.

Some formulas:

1. Normal model, n observations, $\sigma = \sigma_o$ known

   $(\overline{x} - 1.645\ \sigma_o/\sqrt{n},\ \overline{x} + 1.645\ \sigma_o/\sqrt{n})$ is a .90 confidence interval for the mean μ.

2. Normal model, n observations, σ unknown

   $\left(\overline{x} - t_{n-1,.95} \frac{s}{\sqrt{n}},\ \overline{x} + t_{n-1,.95} \frac{s}{\sqrt{n}}\right)$ is a .90 confidence interval for μ,

   where $s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$ and $t_{n-1,.95}$ is the .95 critical value of

   Student's t distribution with n - 1 degrees of freedom.

3. Binomial model, large np (say 10 or more)

   $\left(\hat{p} - 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ is a .95 confidence interval for

   for p.

A more accurate confidence interval for the binomial parameter is given by

$$\frac{n}{n + (1.645)^2} \left[\hat{p} + \frac{(1.645)^2}{2n} \pm 1.645 \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(1.645)^2}{4n^2}}\ \right].$$

For np or n(1-p) small, one should use special graphs prepared for the binomial to obtain confidence limits on p. These are very useful and are given in many textbooks on statistics.

The above formulas may readily be developed. Consider the first case, in which we have a normal model with known variance. We have the pivotal function

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

which is based on the statistic $\overline{X}$, parameter of interest μ, known constants σ, n, and which has a fixed normal distribution. We may write

$$P\left(-1.645 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.645\right) = .90$$

and, rearranging, obtain

$$P\left(\bar{X} - 1.645 \frac{\sigma}{\sqrt{n}} \leqq \mu \leqq \bar{X} + 1.645 \frac{\sigma}{\sqrt{n}}\right) = .90.$$

For the binomial model we have

$$P\left(-1.645 \leqq \frac{\hat{p} - p}{\sqrt{}} \leqq 1.645\right) = .90,$$

with $\sqrt{}$ being the standard deviation of $\hat{p}$. The first result uses $\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$ and the second uses $\sqrt{\frac{p(1 - p)}{n}}$. With these denominators we again need only to rearrange the probability statement so that p is in the middle.

Sample Size Problems:

1. Suppose we wish to estimate the fraction of students on campus who own an automobile. How large a sample is needed to assure us that

$$P(|\hat{p} - p| \leqq .02) = .90?$$

Now it is approximately true that

$$P\left(|\hat{p} - p| \leqq 1.645 \sqrt{\frac{p(1 - p)}{n}}\right) = .90$$

and we shall be conservative (have too large an n) if we simply seek an n such that $1.645 \sqrt{\frac{(.5)(.5)}{n}} = .02$.

2. Suppose we wish to estimate the mean of a normal population (say the average height of the freshmen students). We shall use $\bar{X}$ as our estimate. How large a sample is needed in order that we may assert

$$P(|\bar{X} - \mu| \leqq .1\sigma) = .90?$$

We know that

$$P(|\hat{X} - \mu| \leqq 1.645\ \sigma/\sqrt{n}) = .90$$

and hence we need an n such that $\frac{1.645}{\sqrt{n}} = .1$; i.e., n = 271.

Tests for equality of binomial parameters

Suppose as a result of a survey of freshmen students we find

	Smokes	Doesn't Smoke
Boys	20	180
Girls	15	185

At what level of significance would you say that there is a difference in the proportion of boy and girl smokers? We assume that there are many more than 200 boys and 200 girls in the freshmen class. This is simply a test of equality of two binomial parameters, so the chi-square goodness of fit test is appropriate. We calculate a table of "expected values" or "estimated values" under the assumption of independence. These expected frequencies are

	Smoke	Doesn't Smoke	
Boys	17.5	182.5	200
Girls	17.5	182.5	200
	35	365	400

The statistic we calculate is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(2.5)^2}{17.5} + \frac{(2.5)^2}{17.5} + \frac{(2.5)^2}{182.5} + \frac{(2.5)^2}{182.5}$$

$$= \frac{12.5}{17.5} + \frac{12.5}{182.5} < 1.$$

With what do we compare this value?--a critical value (say .05) of χ^2 with 1 d.f. Since this value is 3.841, we could not say that the proportion of boys who smoke is higher than the proportion of girls.

Problem: Verify that the chi-square goodness of fit test is equivalent to using the normal approximation to compare equality of the parameters p_1 and p_2 in two binomial models.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 6: Contingency Tables

Yesterday we considered tests for equality of binomial parameters using enumeration data which were classified according to two characteristics. Our example considered sex (boys, girls) as one characteristic, the other being smoking trait (smoker, non-smoker). The data were summarized into what is called a contingency table. Let us consider another interesting problem along these lines.

Problem: In order to investigate the effects, if any, of atmospheres A_1, A_2, A_3, and A_4 upon the incidence of lung cancer, a number of mice were placed in pens of each of these atmospheres for one year. After this time their lungs were examined and the presence or absence of the symptoms of lung cancer was noted. The four atmospheres considered were:

A_1: pure or natural air

A_2: high concentration of a germ carrier

A_3: high concentration of a toxin

A_4: a combination of A_2 and A_3

It is interesting to note that the original experiment led to some unexpected difficulties. Some of the penned mice turned cannibal and destroyed much of the hopeful data. This was particularly bad, as a rather small number of mice were involved in the experiment--perhaps consulting an animal husbandry expert would have prevented this difficulty. Another difficulty which developed was that no tumors were found, so other measures of cancer symptoms were required. Nevertheless it was possible to conclude that A_2 appeared harmful, A_3 had no effect, and A_4 was not particularly harmful. This led to speculation that toxic A_3 counteracted the germs carried in A_2.

Let us consider the analysis of a contrived set of such data where 1000 mice were subjected to the four atmospheres.

Response	A_1	A_2	A_3	A_4	Total
Dead	100	140	180	200	620
Live with symptoms	5	19	10	14	48
Live with no symptoms	895	841	810	786	3332
Totals	1000	1000	1000	1000	4000

Questions:

a) What can you say about the effects of the atmospheres upon the survival rates?

b) What can you say about the effect of the atmospheres upon the presence of the symptoms of lung cancer?

The first question is concerned with only survival rates, so we first condense the data to the appropriate response of dead or alive by atmosphere. This is shown in the following table:

Response	A_1	A_2	A_3	A_4	Total
Dead	100	140	180	200	620
Alive	900	860	820	800	3380
Totals	1000	1000	1000	1000	4000

The hypothesis we wish to test is

$$H_0: p_1 = p_2 = p_3 = p_4$$

against the alternative

$$H_1: p_i \neq p_j$$

for some i, j, where p_i is defined as the probability of survival when subjected to atmosphere A_i, i = 1, 2, 3, 4. We may now test for significance using the chi-square test by comparing these observed cell frequencies with the expected cell frequencies. We see by multiplying marginal totals and dividing by the grand total that the expected cell frequencies are 155 for the top row and 845 for the bottom row. Our test statistic is then

$$\chi^2 = \sum_{i=1}^{8} \frac{(0_i - E_i)^2}{E_i} = \frac{(100 - 155)^2}{155} + \dots + \frac{(800 - 845)^2}{845} ,$$

which is clearly quite large and hence significant when compared to percentage points of the chi-square with three degrees of freedom. We conclude then that survival rates differ.

Can we carry the analysis further? Perhaps we would like to compare pairs of atmospheres to see where the differences occur. We can do this in a like manner to the above. However, are these further comparisons prompted by looking at the data or were they hypothesized to be of interest prior to experimentation? We must be careful here to ensure the true level of significance of the tests, but we also most assuredly want to use the experimental results to the fullest degree. There is no universal agreement among statisticians on how to resolve difficulties here.

Consider an experiment consisting of K treatments, say $M_1, M_2, \dots, M_K$, with independent random samples of n_i, $i = 1, \dots, K$, observations made on each treatment. If we wish to compare, say, treatment M_2 with treatment M_K we would calculate the test statistic

$$\frac{|\overline{X}_2 - \overline{X}_K|}{\hat{\sigma}\left(\frac{1}{n_2} + \frac{1}{n_K}\right)^{\frac{1}{2}}} ,$$

where $\hat{\sigma}$ denotes the standard error estimated by all the data. Using normal tables we would reject the hypothesis of equal effects of the two treatments at the .05 level of significance if this test statistic equals or exceeds 1.96. This test would not be valid if we knew (from looking at the data) that $\overline{X}_2$ and $\overline{X}_K$ were the largest and smallest means in the set. Another problem is that with successive comparisons the overall error rate is compounded. A number of methods are available here. Let's look at the more popular.

Fisher's Least Significant Difference (LSD) would declare the i^{th} and j^{th} means as being significantly different at the .05 level of significance if

$$|\overline{X}_i - \overline{X}_j| \geq 1.96\ \hat{\sigma}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{\frac{1}{2}}$$

$$\geq \ell(n_i, n_j, \alpha)\ \hat{\sigma}.$$

This procedure ensures that the error rate for any such comparison is in fact α.

Tukey's method requires special tables and declares significant differences if

$$|\overline{X}_i - \overline{X}_j| \geq \text{Tukey}(n_i, n_j, \alpha)\ \hat{\sigma}.$$

This procedure ensures that the experiment wide error rate is at most α. That is, no matter how many pairwise comparisons are made, the overall error rate is controlled at α. If the number of treatments is large, Tukey's test would not be able to detect pairwise differences, whereas Fisher's test is not affected by the number of treatments.

There are other popular procedures such as Duncan's New Multiple Range Test and the Scheffé test. The latter is much like Tukey's test, but all tests of linear combinations of means may be handled. All of these procedures are techniques of dealing with the dilemma that if the experiment wide error rate is held down, then we pay the price of not being able to detect true differences; on the other hand, if we control the individual comparison error rate, the experiment wide error rate is inflated. My preference is to be able to detect all differences.

Extra Binomial Problems:

1. How many apple trees should I plant if I wish to have probability of .9 that there will be 100 or more living trees at the end of 3 years, given that the probability a planted tree is living 3 years later is .7? Answer: 153.

This is the same problem that a manufacturer confronts when he has a machine that makes a defective part .3 of the time and he wishes to make at least 100 good parts with probability of .9.

2. How many freshmen students should we accept if we have space for 1000 students and the probability that an accepted student shows up is .7? Let us say we wish to have probability of .8 that our space limit will not be exceeded.

3. Suppose Glen has 12 graduate assistantships to fill and he is willing to assume the probability of acceptance is .6. Let X be the number of acceptances he receives from N offers. How large can N be to assure that $P(X \leq 12) = .9$?

Answer: From tables of the binomial we have

$$P(X \leq 12) = \begin{cases} .874, & n = 17 \\ .934, & n = 16 \end{cases}$$

4. Suppose the Army needs 10,000 men next month and the rejection rate is .5. How many men should the Army have examined in order to have probability of .9 that at least 10,000 will pass?

Answer: Using the normal approximation, $n = 20{,}149$.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 7: Sampling Problems

As we become more interested in our environment, we would like to have answers to the following questions:

1. What is the average thickness of the topsoil in Phelps County, Missouri?

2. What is the average level of CO_2 in the air in Gary, Indiana?

3. How much non-degradable garbage does a person in New York City produce each day on the average?

4. Are the anti-pollution devices on the automobiles in Los Angeles working properly? What fraction of them is working properly?

5. The State of Pennsylvania wished to learn how much it would cost to light adequately all of the accident locations in the state that had a "high" night to day ratio and also had a high number of night accidents. The accident reports for the past year were available on tape for the state. These reports gave location, severity of accident, time of accident, etc. Discuss how one could have obtained the desired information.

6. How would you determine whether the ratio of smokers among students differs from the ratio of smokers among teachers?

7. How would you determine whether the ratio of G.I.'s in Viet Nam who smoke pot differs from the ratio of university students who smoke pot?

8. What sort of a poll should Nixon take to determine if he is likely to win California in 1972?

9. How would you estimate the number of teenagers in New York who are on "hard drugs?"

Let's discuss the sample survey techniques needed to determine answers to some of these questions.

To determine the average thickness of topsoil in Phelps County, the best thing to do, I suppose, is to call the county agent or some other local expert. But let's approach the problem from a statistical point of view. The data to be collected will be readings of topsoil depth, say in inches, taken at various test sites. The problem of interest is how to select these test sites or sampling points. Reasonable approaches to this problem are:

A. Mark off the county into a very fine grid system so that a set of points could randomly be selected throughout the county. The idea here is to allow each point in the county to have the same probability of being selected as a sample point.

B. Divide the county into its distinct land types and then sample within these distinct areas. This procedure ensures that each land type will be sampled.

C. Mark off the county into a grid system. Then sample each intersection point. This method gives a ready scheme for collecting the data.

These three techniques are known, respectively, as simple random sampling, stratified sampling, and systematic sampling. Let's look at some statistical aspects of these sampling techniques in light of our topsoil problem. If we look at the problem from an ecology point of view, we are interested then in estimating the amount of topsoil available today; then, in, say, twenty years we will want to re-estimate to see whether we're "losing ground." Define the quantity of interest as the total volume of topsoil in Phelps County,

$$T = \int \int f(x_1,x_2)\ dx_1\ dx_2 ,$$

where $f(x_1,x_2)$ is a function representing the depth of topsoil at the point (x_1,x_2) and, of course, the region of integration is that of Phelps County.

A. Simple Random Sampling

We assume a simple random sample $x_1, \ldots, x_n$. The x's are independent measurements of topsoil depth taken randomly from Phelps County. The estimator of total volume of topsoil in Phelps County is then

$$\hat{T} = \left\{ \sum_{i=1}^{n} x_i/n \right\} \text{(Area of Phelps County)}.$$

The variance of this estimator is then

325

$$\text{var } \hat{T} = (\text{Area of Phelps County})^2 \; \frac{\text{var } (x_i)}{n},$$

which would itself be estimated by

$$\widehat{\text{var } \hat{T}} = (\text{Area of Phelps County})^2 \; \frac{s_x^2}{n},$$

where s_x^2 is the usual estimate of a population variance. With these, the sample size required may be determined using the results from Lecture 5.

B. Stratified Sampling

Stratified sampling requires us to divide the county into regions $A_1, A_2, \ldots, A_H$ which have uniform amounts of topsoil. For example, A_1 may be mountainous regions, A_2 city regions, A_3 fertile valleys, etc. Ideally we would like to know the region of zero topsoil, the region with one inch of topsoil, etc. In many problems these regions or strata are well-defined from physical considerations. Such problems naturally suggest stratified sampling, but this technique is also commonly used in situations where data are costly and precision is required. Even with well-defined strata, there is the problem of how many samples should be taken from each strata. This depends on the variability within the strata and also the area of the region. We will also use the symbol A_i to represent the area of region A_i. Data for a stratified sampling experiment would appear as:

Area of region:	A_1	A_2	$\cdots$	A_H
Sample size:	n_1	n_2	$\cdots$	n_H
Observe:	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_H$

where $\bar{x}_j$ is the mean of the n_j independent observations recorded on the j strata, j = 1, 2, ..., H. The estimator of total volume of topsoil is then

$$\hat{T} = \sum_{j=1}^{H} \bar{x}_j \, A_j$$

with an estimated variance of

$$\widehat{\mathrm{var}\ \hat{T}} = \sum_{j=1}^{H} A_j^{2}\ \widehat{\mathrm{var}\ (\bar{x}_j)}$$

$$= \sum_{j=1}^{H} A_j^{2}\ \frac{S_j^{2}}{n_j}$$

C. Systematic Sampling

Systematic sampling is not used very often. I feel that systematic sampling data may be analyzed just like a simple random sample, with the results being on the conservative side; there is, of course, no probability model involved.

Using these sampling schemes we can now present the estimate of total volume of topsoil in Phelps County in terms of a confidence interval or as a point estimate along with its standard error.

Lecture 8: Location and Scale Parameter Families

The following families of densities have the common property that they are location and scale parameter densities:

1. Normal

$$f(x;a,b) = \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{1}{2}\left(\frac{x-a}{b}\right)^2}$$

2. Laplace

$$f(x;a,b) = \frac{1}{2b} e^{-\left|\frac{x-a}{b}\right|}$$

3. Two-parameter exponential

$$f(x;a,b) = \frac{1}{b} e^{-\left(\frac{x-a}{b}\right)} \quad \text{for } x \geqq a$$

4. Logistic

$$f(x;a,b) = \frac{1}{b} \frac{e^{-\left(\frac{x-a}{b}\right)}}{[1 + e^{-\left(\frac{x-a}{b}\right)}]^2}$$

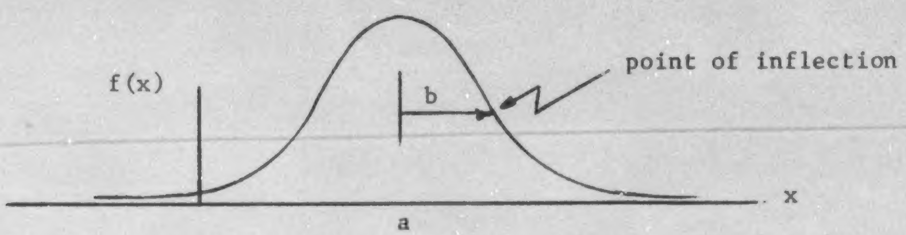
5. Cauchy

$$f(x;a,b) = \frac{1}{\pi b[1 + \left(\frac{x-a}{b}\right)^2]}$$

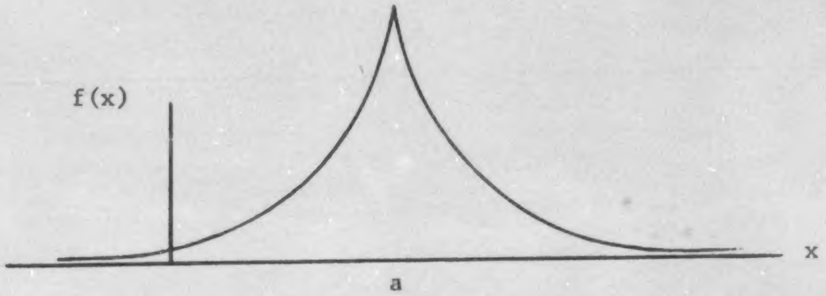
6. Type 1 extreme value for maximums

$$f(x;a,b) = \frac{1}{b} e^{-[e^{-\left(\frac{x-a}{b}\right)} - \frac{x-a}{b}]}$$

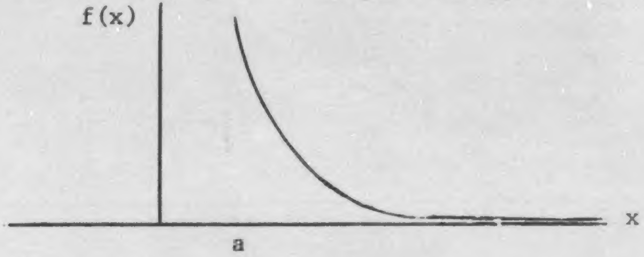
The normal distribution is familiar to all of us. It is a two-parameter model with the location parameter being the mean and the scale parameter being the standard deviation. In all the models above these parameters appear as a and b, respectively. Of course, if we know a and b the density is completely specified and might be graphed as the well known bell-shaped curve.



The Laplace distribution also depends on the location parameter a and scale parameter b. The density is symmetric about a and has the form



The two-parameter exponential is the only one of the above densities in which the random variable has a restricted range. It appears as



The logistic distribution looks very much like the normal; in fact, statistical tests for discriminating between these two models are not at all effective. The logistic model does, however, have a very nice cumulative distribution function which is given by

$$F(x) = [1 + e^{-\left(\frac{x-a}{b}\right)}]^{-1}.$$

Our use of the logistic model is not to be confused with the logistic growth curve which is often used to determine drug dose levels required to ensure some cure rate.

The Type 1 extreme value model also has a nice cumulative distribution function which is given by

$$F(x) = e^{-\left[e^{-\left(\frac{x-a}{b}\right)}\right]}.$$

We shall return to this model shortly.

Now we all know how to obtain confidence limits for a and b (more commonly denoted μ and σ) if we have the normal model. In this case we make use of the familiar pivotal functions

$$\sqrt{n}\,\frac{(\bar{X}-\mu)}{\sigma} \sim N(0,1),$$

$$\sqrt{n}\,\frac{(\bar{X}-\mu)}{S} \sim t_{n-1},$$

and $$(n-1)\,\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

In the normal case the result which allows us to construct confidence limits for the parameters a and b is the fact that the densities of

$$\sqrt{n}\,\frac{(\hat{a}-a)}{b},$$

$$\sqrt{n}\,\frac{(\hat{a}-a)}{\hat{b}},$$

and $$\frac{\hat{b}}{b}$$

do not depend upon a or b. Would this be true for the other densities where $\hat{a}$ and $\hat{b}$ are the maximum likelihood estimates for a and b? Yes; hence these same pivotal functions may be used to set confidence limits on the parameters or to test hypotheses about them just as in the normal case.

If the tables for the particular model and desired sample size n are not available, then it is fairly easy to simulate with the computer, say, 2000 sets of samples of size n and thus generate the needed tables. This would be a reasonable project for a student who wished to have some simulation work to do. The biggest problem for some of these (logistic, Cauchy, Type 1 extreme value) is to find the maximum likelihood estimates. They require one to solve a system of nonlinear equations, but there always exists a unique solution for the logistic and Type 1 extreme value cases and we believe that with $n > 2$ there will (with probability 1) always be a unique solution for the Cauchy case. The procedure is outlined below for the Laplace distribution.

1. Generate sample of size n from uniform (0,1) say $u_1, u_2, \ldots, u_n$.

2. Let $x_i = \begin{cases} \ln 2u_i \text{ if } u_i \leq 1/2 \\ -\ln 2(1 - u_i) \text{ if } u_i > 1/2 \end{cases}$

3. From the x_i's calculate $\hat{a} = \text{Med } x_i$ and $\hat{b} = \frac{1}{n}\sum |x_i - \hat{a}|$, the maximum likelihood estimates of a and b.

4. "Score" $\hat{a}$, $\hat{a}/\hat{b}$, and $\hat{b}$ and return to Step 1 if this is not the 2000th trial.

It is also true that $F(X;\hat{a},\hat{b})$ has a density that does not depend upon a or b and hence one can always obtain confidence limits for the distribution function. This fact is used to enable one to set confidence limits on the reliability. The reliability for "time" x is the probability that $X \geq x$, i.e., $1 - F(x;a,b)$. In most cases one must use the general method to obtain these confidence limits for the reliability and do this by simulation.

These facts open the way for one to do with any location and scale parameter model all the work that one might do with the normal model.

As an example, let $X_1, X_2, \ldots, X_n$ be the maximum yearly flood levels for a certain stream for the past n years. Let X_M be the maximum flood level for this stream in the next M years. Then we'd like to have an upper .99 confidence limit on X_M. It is often reasonable to assume that our observations X_i have a Type 1 extreme value distribution; if this is the case, then it is easily seen that $\frac{X_M - \hat{a}}{\hat{b}}$ has a density that does not depend upon a or b and hence it may be used to set confidence limits on X_M. This would be very useful to the people designing projects along rivers.

Problem: Let $X_1^*, X_2^*, \ldots, X_K^*$ be the Type 1 extreme value random variables. Show that if $X_M = \mathrm{Max}\{X_1^*, \ldots, X_K^*\}$, then X_M also has a Type 1 extreme value density with parameters a + b ln K and b, where a and b are the location and scale parameters for the X_i^*'s.

Problem: Show that

$$\frac{X_M - \hat{a} - \hat{b} \ln K}{\hat{b}}$$

has a density that does not depend upon a, b, or K. This is the pivot function which we used to obtain confidence limits for X_K so that we would not need to use K in the tables.

Let us consider these problems in the framework of the previous example. That is, we have the random sample $X_1, \ldots, X_n$ of maximum yearly flood levels for the past n years. Let $X_1^*, \ldots, X_K^*$ be the maximum flood levels for the next K years. We now require the distribution of $X_M = \mathrm{Max}\{X_1^*, \ldots, X_K^*\}$. We may write

$$\begin{aligned} F_{X_M}(Z) &= P(X_M \leqq Z) \\ &= P(X_1^* \leqq Z, \ldots, X_K^* \leqq Z) \\ &= [P(X^* \leqq Z)]^K, \text{ by independence} \\ &= [F_{X^*}(Z)]^K \\ &= [e^{-[e^{-\left(\frac{Z-a}{b}\right)}]}]^K \\ &= e^{-[e^{-\left(\frac{Z - (a + b \ln K)}{b}\right)}]} . \end{aligned}$$

This solves the first problem. Note that the effect here is just an upward shift in the location parameter.

For the second problem we have the variate X_M is Type 1 extreme value with parameters a + b ln K and b. Consider the standardized variable

$$\frac{X_M - (a + b \ln K)}{b} = \frac{X_M - a}{b} - \ln K$$

having a location parameter of zero and a unit scale parameter, and hence a density independent of a, b, and K. Also, we know that $b/\hat{b}$ has a density independent of b, so that

$$\frac{X_M - a}{b} \frac{b}{\hat{b}} - \ln K$$

has density independent of a, b, and K. Now subtracting

$$\frac{\hat{a} - a}{\hat{b}} ,$$

which is independent of a and b, we see that

$$\frac{X_M - \hat{a}}{\hat{b}} - \ln K$$

has density independent of a, b, and K. We may, therefore, readily generate percentage points of this random variable.

Lecture 9: An Application of Markov Chain Theory

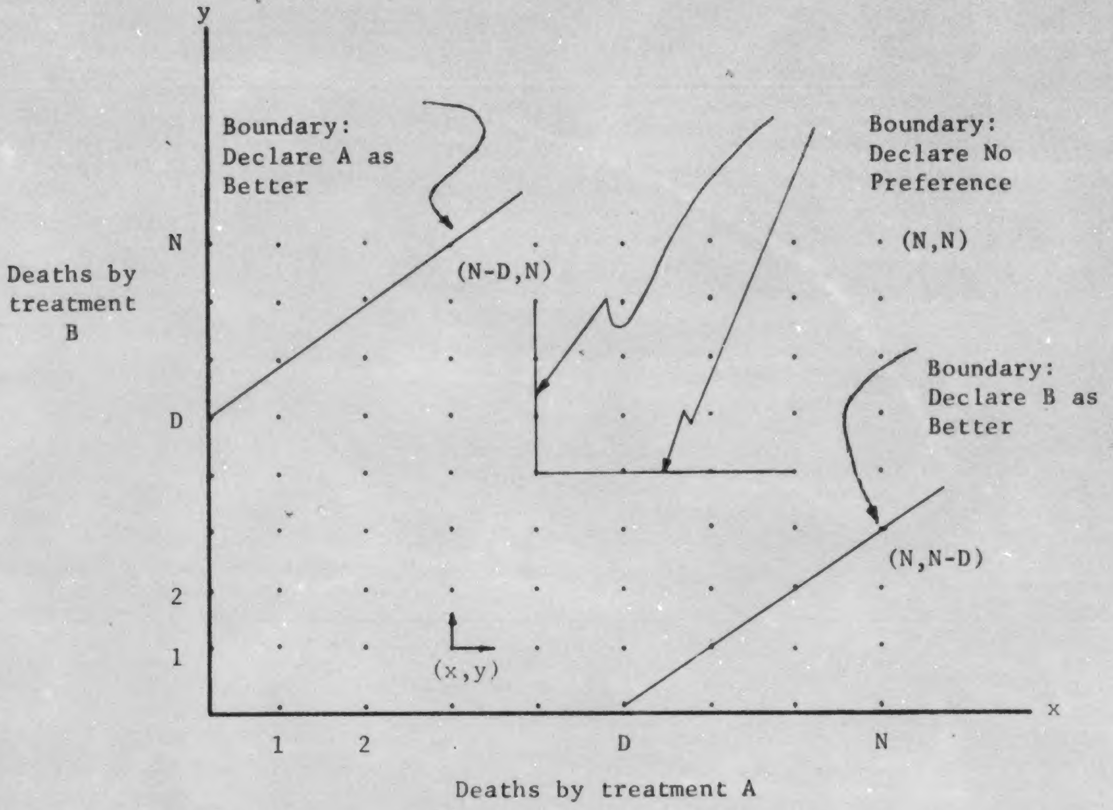
While consulting with an engineer about which of two motors he should recommend for use, he informed me that he had been running five of each kind for a long time and that one of brand A had failed while three of brand B had failed. With just this amount of information and no further assumptions, the engineer wanted to know whether he could terminate the experiment and be able to draw some conclusions. In particular, could he declare brand A as the better motor brand and, if so, at what level of significance? To establish the level of significance we require that the probability of an outcome be as bad as or worse than that observed. For this problem we want the probability, given that four motors failed when the ten were tested, that exactly three of the motors which failed were of one type or that all four were of one type, assuming both types were equally good. The complement of this would be the situation that, of the four failures, two of each brand failed. This probability may readily be evaluated using the hypergeometric model

$$\frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}} = \frac{100}{210}.$$

Hence, the desired probability is 11/21, not a very desirable level of significance. This really isn't a very good approach to the problem of comparing two brands of motors. The sample sizes are quite small; also, if possible, we should use time-to-failure data. The latter would allow us to use a more appropriate model. We also need some stopping rule for terminating the experiment so that conclusions can be drawn.

Another problem similar to the above is that of comparing two treatments, say A and B, when it takes a long time to observe the success or failure of the treatments. Let us assume that the two treatments are drugs for a particular illness. We give treatment A to N people and treatment B to N people, and we shall terminate the experiment whenever the difference in reported responses equals D. What level of significance will be reported when we stop the experiment? To be a bit more specific, let the reported responses be deaths, so that it is desirable to have the response time as long as possible and also desirable to terminate the experiment as early as possible and then give all the living patients the better drug.

The stopping rule for our experiment is to terminate testing if $|X - Y| \geqq D$ or if we see that $|X - Y|$ can never reach D, where X, Y are the number of deaths reported for treatment A, B, respectively. (Here, we are not using time to failure, so perhaps our testing technique could be improved.) This stopping rule has the feature that the experiment will be terminated early if one drug is greatly better than the other. The following figure will be convenient:


y
Boundary: Declare A as Better
Boundary: Declare No Preference
N
(N,N)
Deaths by treatment B
(N-D,N)
Boundary: Declare B as Better
D
(N,N-D)
2
1
(x,y)
x
1
2
D
N
Deaths by treatment A

Each point or coordinate (x,y) represents a "state" the experiment is in, i.e., x deaths for drug A and y deaths for drug B at some time. We note that our stopping rule does not allow all states to occur, as decision boundaries block off the upper right corner of states. At time zero or at the beginning of the experiment we start in state (0,0) and move one step upward or one step to the right depending on whether the next death recorded is for a patient under treatment B or A, respectively. This movement continues as each death is recorded until one of the three boundaries is reached and the testing is terminated.

We would now like to look at the above testing procedure probabilistically. Let P(x,y) be the probability of ultimate capture on the boundary for which we declare B to be the better of the two treatments. Let $P_R(x,y)$ and $P_U(x,y)$ be the probability that we move from state (x,y) one step to the right or upward, respectively. Then for points inside the boundaries we have

$$P_R(x,y) = \frac{N - x}{2N - x - y}$$

and

$$P_U(x,y) = \frac{N - y}{2N - x - y} .$$

These transition probabilities are based on the assumption (or null hypothesis) that treatments A and B are equally effective. We may now write the difference equation

$$\begin{aligned} P(x,y) &= P(\text{ultimate capture on the boundary where B is declared the better}) \\ &= P_R(x,y)\, P(x + 1, y) + P_U(x,y)\, P(x, y + 1), \end{aligned}$$

which has the boundary conditions

P(x,y) = 1 for points on the boundary for which treatment B is declared the better,

P(x,y) = 0 for the other boundary points where the experiment is terminated.

We would now like to find the solution to this difference equation. In particular, we want to find P(0,0). This is not a very easy problem, but it can be solved on the computer by tracing out all possible paths and associated probabilities in a backward fashion from all points on the B boundary to the origin.

Similar probabilities may be developed for ultimate capture on the other two boundaries. All of these probabilities are dependent upon values of N and D. Tables of the probability of declaring no difference in the two treatments under the assumption of no difference may be built up as functions of N and D to aid an experimenter in selecting an appropriate testing procedure. One would also require such tables for assuming the existence of differences between the treatments; here a distribution of time to failure might be imposed. Antle, Dumoncearx, and Klimko have worked on various aspects of these problems.

Another method for comparing two drugs with clinical trials that has received attention in the literature is what is commonly called the "play-the-winner rule." The first patient is randomly assigned drug A or B. If the treatment proves successful, the next patient receives the same treatment; if the treatment is unsuccessful, the next patient receives the other treatment. A typical sequence of trials may look like the following:

Treatment A:	SSF		SF		...	
Treatment B:		SF		SSF		...

Here the first patient receives drug A, which proves successful; the second patient receives drug A with success; the third patient receives drug A without success, so the fourth patient receives drug B, etc. The object of this strategy is to give the better drug to the greatest number of patients. We need, of course, a stopping rule for this testing procedure. Such a rule must ensure that a sufficient number of patients have received each treatment, for the play-the-winner rule as it stands does not even ensure that the better drug will ever be tested. Such considerations lead to the numerous modifications of this rule which can be found in the literature.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 10: Elementary Bayesian Statistics

Suppose you are comparing two drugs by giving them to pairs of patients as the patients present themselves at a clinic and you record a 1 if the drug is beneficial and a 0 if it is not. At some point in the experiment you wish to terminate the experiment and declare one of the drugs to be the better or declare that there is little or no difference. Consider the following sequence of data:

	1	2	3	4	5	6	7
Drug A	0	0	1	0	1	0	0
Drug B	1	1	0	1	1	1	0

Can the experiment be terminated and a decision made? If not, how many trials are required? Clearly, at the end of seven trials drug B appears to be the better and if only two more patients were ever to be treated I would use drug B. However, if many, many patients were yet to be treated I would probably do more testing, as I doubt that I have very conclusive evidence at this stage that drug B is indeed the better. We shall look at this experiment and its analysis from a Bayesian point of view. First, however, let us consider a simpler problem from a Bayesian point of view, that of making inferences about a single binomial parameter.

Single Binomial Problem

Let us consider a problem similar to the above. Assume that there is a standard drug for a given illness which is known to be beneficial for sixty per cent of the patients. That is,

$$p_o = P(\text{standard drug is beneficial}) = 0.60.$$

A new drug is developed for this illness, and we wish to determine whether it is indeed better than the standard drug. That is, determine whether $p > p_o$, where the unknown p is defined as

$$p = P(\text{new drug is beneficial}).$$

The test is to be based on the results of n independent Bernoulli trials. Let x be the number of successes; then x is a binomial variate with parameters n and p.

As the parameter p is unknown, suppose we assign the uniform (0,1) density as our prior density for p. That is, treat p as a random variable with density

$$f(p) = 1, \quad 0 \leqq p \leqq 1.$$

Then after n trials with x successes we have for the conditional density of p, given x,

$$f(p|x) = \frac{f(x,p)}{f(x)} = \frac{\binom{n}{x} p^{x}(1-p)^{n-1}}{\int_0^1 \binom{n}{x} p^{x}(1-p)^{n-x}\, dp}, \quad 0 \leqq p \leqq 1.$$

This conditional density is called the posterior density. We may simplify the expression by using the definition of the Beta function, B:

$$B(m,n) = \int_0^1 z^{m-1}\,(1-z)^{n-1}\, dz = \frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)}, \quad m, n > 0$$

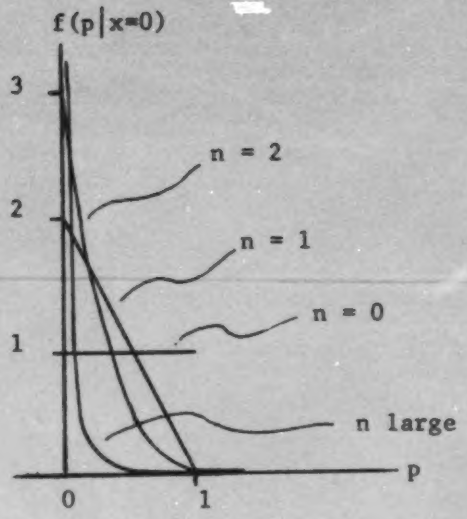
Then for the posterior density of p we have:

$$f(p|x) = (n+1)\binom{n}{x} p^{x}\,(1-p)^{n-x}, \quad 0 \leqq p \leqq 1$$

One can readily verify that this is a valid density function. We see immediately that $f(p|x) \geqq 0$ for $0 \leqq p \leqq 1$ and we use the Beta definition above to show that

$$\int_0^1 f(p|x)\, dp = 1.$$

This posterior density of p is a family of curves whose form depends on the non-negative integers n and x ($x \leqq n$). We see, for example:



Let's return now to the problem of determining whether or not the new drug is better than the standard drug. We can determine probabilistically whether $p > p_o$ now that we have the posterior density for p. We have

$$
\begin{aligned}
P(p > p_o|n,x) &= \int_{p_o}^{1} f(p|x)\,dp \\
&= 1 - \int_{0}^{p_o} f(p|x)\,dp \\
&= 1 - (n+1)\binom{n}{x}\int_{0}^{p_o} p^x(1-p)^{n-x}\,dp.
\end{aligned}
$$

Notice that the integration may be simplified using the relation

$$
\begin{aligned}
\int p^x(1-p)^{n-x}\,dp &= \int p^x \sum_{i=0}^{n-x} (-1)^i \binom{n-x}{i} p^i\,dp \\
&= \sum_{i=0}^{n-x} (-1)^i \binom{n-x}{i} \int p^{x+i}\,dp \\
&= \sum_{i=0}^{n-x} (-1)^i \binom{n-x}{i} \frac{p^{x+1+i}}{x+1+i} + C.
\end{aligned}
$$

For our problem we have $p_o = 0.60$. The trial results for drug A from page give

	Trial							
	0	1	2	3	4	5	6	7
Drug A score	0	0	0	1	0	1	0	0
$P(p > 0.60 \mid n,x)$	.4	.16	.064	.179	.087	.179	etc.	

If we had observed, for example, that the drug was not beneficial in the first three trials then we would have

$$P(p > 0.60 \mid n = 3,\ x = 0) = 0.0256.$$

Is this probability sufficiently low to declare that the drug is not more effective than the standard drug? Care should be taken on interpreting these results, as the uniform model is perhaps not too good a prior density to use. This choice of prior is a personal or subjective, not an objective, choice.

In our example we have $n = 7$, $x = 2$. What value should be given as a point estimate of p? The classical approach would be to give the maximum likelihood estimate $\hat{p} = \frac{x}{n} = \frac{2}{7}$ by treating this as a binomial problem. The Bayesian might report the most probable value which would be the mode of the posterior. This too is $\hat{p} = \frac{x}{n}$ for our choice of prior. If he were to report the minimum mean square error estimate, he would give $E(p \mid x) = \frac{x+1}{n+2} = \frac{1}{3}$. If an interval estimate were required, the Bayesian would report p_L which satisfies $p(p \geqq p_L \mid x) = \gamma$. This involves solving

$$\gamma = \int_{p_L}^{1} f(p \mid x)\, dp$$

for p_L. Such a p_L would replace the classical lower confidence limit for p.

Two Binomial Problem

Let us now return to the two drug problem as outlined on the first page of this section. Recall that we want to determine whether two drugs differ, basing our decision on n clinical trials for each drug. Notationally we have, for $i = 1, 2$,

$$p_i = P(\text{drug } i \text{ is beneficial})$$

and

$$x_i = \text{the number of successes of the } i^{th} \text{ drug in n trials.}$$

For the Bayesian approach, to compare the two drugs we would like to be able to calculate

$$p = P(p_1 > p_2 | x_1, x_2).$$

Starting with the prior density of the unknown parameters p_1 and p_2 as being uniform over the unit square, i.e.,

$$f(p_1, p_2) = 1, \quad 0 \leq p_i \leq 1, \quad i = 1, 2,$$

we have as the joint posterior density for p_1, p_2

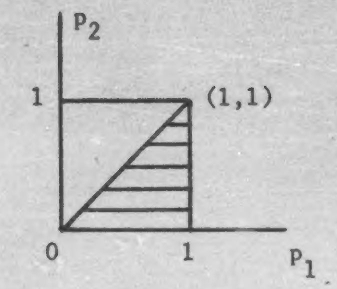
$$f(p_1, p_2 | x_1, x_2) = (n + 1)^2 \binom{n}{x_1}\binom{n}{x_2} p_1^{x_1} (1 - p_1)^{n-x_1} p_2^{x_2} (1 - p_2)^{n-x_2},$$

$$0 \leq p_i \leq 1, \quad i = 1, 2.$$

To find the probability of interest we have

$$p = P(p_1 > p_2 | x_1, x_2) = \int\int f(p_1, p_2 | x_1, x_2)\, dp_1\, dp_2,$$

where the integration is over the region $p_1 > p_2$ depicted as the shaded region



and $f(p_1, p_2 | x_1, x_2)$ is a surface over this region. Thus:

$$n = 1,\ x_1 = 0,\ x_2 = 1 \Longrightarrow p = 4 \int_0^1 \int_{p_2}^1 (1 - p_1)\, p_2\, dp_1\, dp_2 = 1/16$$

$$n = 2,\ x_1 = 0,\ x_2 = 2 \Longrightarrow p = 9 \int_0^1 \int_{p_2}^1 (1 - p_1)^2\, p_2^2\, dp_1\, dp_2 = 1/20$$

etc.

These probabilities are to be used to decide whether the drugs differ. Of course, these probabilities are highly dependent on our uniform prior. Perhaps some other prior would be more appropriate. We would like to develop a stopping rule for the experiment, based on probabilities such as above. The stopping rule should be based on the seriousness of the decision and on how much use will be made of the decision (will there be 10 future patients or 10,000?).

Lecture 11: Regression Analysis, an Example

I like to introduce a class project on regression analysis by posing the following questions:

> How could I use a tape measure to obtain the weight of a male student? Which measurements--waist size, height, hat size, etc.--should I use in order to predict his weight? Could I devise a formula and thereby predict his weight better than the man at the fair who guesses within three pounds or gives a twenty-cent prize (the fee is a quarter, of course)?

In order to begin this project we must first decide which measurements are important. Obviously we could take many measurements on an individual--the more measurements, the better the prediction, but the more unwieldy the model. Let's see how well we can predict weight W with measurements H the height, B the waist (belt) size, S the shoe size, and C the collar size. We now need an equation or model relating weight to the other measurements. Consider the model

$$W = B_0 + B_1 H + B_2 B + B_3 S + B_4 C + e,$$

where B_0, B_1, B_2, B_3, and B_4 are parameters to be determined and e represents the error. Is this a reasonable model? Obviously not--so let's add the quadratic terms

$$B_5 H^2 + B_6 B^2 + B_7 S^2 + B_8 C^2,$$

and, since we suspect that weight is somehow related to volume, let's also add

$$B_9 HS^2 + B_{10} HB^2 + B_{11} HC^2.$$

The complete model we want to consider is then

$$W = B_0 + B_1 H + B_2 B + B_3 S + B_4 C + B_5 H^2 + B_6 B^2 + B_7 S^2$$
$$+ B_8 C^2 + B_9 HS^2 + B_{10} HB^2 + B_{11} HC^2 + e,$$

with the twelve parameters B_i, $i = 0, 1, \ldots, 11$, to be determined. The variables (or combination of variables) associated with each parameter were included in the model as they were felt to be useful predictors, prompted perhaps by physical considerations of the problem. Surely some of the variables in our model are unnecessary. There are a number of statistical methods for determining and hence removing the unnecessary variables. The procedure which I recommend is the stepwise procedure.

Let's look briefly at the stepwise procedure. As the name implies, the stepwise procedure builds up a model one variable at a time. To determine the first variable to enter, correlations between the dependent variable, weight, and the eleven independent variables are calculated. The variable which shows the highest correlation is first to enter the model. The second variable to enter is chosen by selecting that one which has the highest partial correlation. That is, the first step is repeated by looking at all correlations, correcting for a variable's being in the model. Now each time a new variable is brought into the model the procedure checks to see whether you may step down, i.e., whether a term already in the model may be removed. This procedure is continued until the addition of a new variable does not significantly improve the model.

Returning to our example problem, data collected in class appeared as:

Shoe Size	Waist Size	Collar Size	Height	Weight
9	34	15.5	68	160
10	32	15.5	70	160
10.5	31	16	71	150
7.5	29	14.5	68	120
8	34	16	68	175
10.5	34	15.5	76	190
12	38	16.5	73.5	205
12	34	17.5	75.5	215
11	36	16.5	73	185
9.5	32	15.5	72	170

In actuality, data on twenty students were used. It is interesting to try to guess the importance of the variables in the model. The first variable to enter the model is almost always anticipated to be height. Let's look at some output from the stepwise analysis. The single-variable model appeared as

$$W = -298.0 + 29.7\ C.$$

It is rather surprising to find collar size to be the singly most important predictor (determined with this set of data, of course). The succeeding models were built up as

$$W = -294.0 + 18.3\ C + 5.3\ B$$

$$W = -359.0 + 13.6\ C + 5.0\ B + 2.1\ H$$

and so on until no significant improvements were made in the model by introducing more terms.

Lecture 12: Analysis of Variance

When I first started studying statistics, the analysis of variance was quite popular and well defined. Analysis of variance is easy to describe from one point of view. From a mathematical statistics point of view, analysis of variance is presented very nicely; in practice, there is a process of communication which builds up. This cannot be found in a book. The experimenter comes to the statistician with data, not with a model. The data must be analyzed and inferences made. Analysis of variance in practice is then an art with a thread of rigor to it. We try to:

(a) partition total variability in a "meaningful" way

(b) think about experimental errors and make intelligent decisions about which terms are indeed measures of error

(c) identify individual degrees of freedom (DF)

In practice, concepts and computations are not always together in these regards.

In analysis of variance we have n observations, say $x_1, \ldots, x_n$. An overall measure of the variability of the data is the total sum of squares $\sum_{i=1}^{n} (x_i - \bar{x})^2$. We want to partition this in a meaningful way--a way which we can interpret. It might be pointed out here that two statisticians may give different answers to a problem because they find different things to be meaningful.

One way to partition the total sum of squares is to use an orthogonal transformation. As an example, consider the data x_1, x_2, and x_3, and make the orthogonal transformation

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} .$$

This transformation amounts to a rotation of axes and is distance preserving. That is,

$$z_1^2 + z_2^2 + z_3^2 = x_1^2 + x_2^2 + x_3^2,$$

with

$$z_1 = \sqrt{3}\,\overline{x}$$

$$z_2^2 + z_3^2 = \sum x_i^2 - 3\overline{x}^2$$

$$= \text{total sum of squares.}$$

The rows of the transformation matrix are orthogonal contrasts, and

$$\sum (x_i - \overline{x})^2 = z_2^2 + z_3^2.$$

We have orthogonally partitioned the total sum of squares with two degrees of freedom into two portions, each having a single degree of freedom. Generally, we should think of analysis of variance as a partitioning of the total sum of squares with n - 1 degrees of freedom into n - 1 individual sums of squares, each having a single degree of freedom. Textbook presentations of analysis of variance, on the other hand, usually group the individual degrees of freedom.

Let us consider the following randomized block experiment. We have two classes (blocks) of three students each, and three different textbooks are used in a given class. We want to compare the three textbooks (and perhaps the two classes). In each class we assign the textbook at random to the students (or vice versa, if you prefer). When the class is over, a final score in the course is recorded as below for each student.

		Textbook		
		1	2	3
Class	1	70	83	91
	2	76	78	65

We could proceed to analyze these data with the textbook approach for a randomized block experiment; however, let's perform the analysis by partitioning the five degrees of freedom for the total sum of squares into five single degree of freedom orthogonal contrasts.

Table of Contrasts

Contrast	Score						
	70	83	91	76	78	65	Divisor
1	1	1	1	-1	-1	-1	$\sqrt{6}$
2	1	-1	0	1	-1	0	$\sqrt{4}$
3	1	1	-2	1	1	-2	$\sqrt{12}$
4	1	-1	0	-1	1	0	$\sqrt{4}$
5	1	1	-2	-1	-1	2	$\sqrt{12}$

These contrasts were written down to be meaningful. Let's look at their meaning:

1. Comparison of class 1 vs. class 2
2. Comparison of book 1 vs. book 2
3. Comparison of the average of books 1 and 2 vs. book 3

We see that contrasts 4 and 5 are measures of interaction, that is, measures of the extent to which the books produce different effects in the two classes. Specifically:

4. Comparison of (book 1 vs. book 2 in class 1) vs. (book 1 vs book 2 in class 2)
5. Comparison of (average of books 1 and 2 vs. book 3 in class 1) vs. (average of books 1 and 2 vs. book 3 in class 2) or (contrast 3 in class 1) vs. (contrast 3 in class 2)

Let's code the data to help with the numerical work. Subtracting 75 from each of the six observations, we obtain the coded data

							Totals
x:	-5	8	16	1	3	-10	13
x^2:	25	64	256	1	9	100	455

and

$$\sum (x_i - \overline{x})^2 = 455 - 13^2/6$$

$$= 426.83.$$

Using the coded data we now calculate the sum of squares of each contrast.

$$z_1^2 = [1(-5) + 1(8) + 1(16) - 1(1) - 1(3) - 1(-10)]^2/6$$

$$= 104.17$$

$$z_2^2 = [-5 - 8 + 1 - 3]^2/4$$

$$= 56.25$$

$$z_3^2 = 2.08$$

$$z_4^2 = 30.25$$

$$z_5^2 = 234.08$$

These calculations are then summarized in the analysis of variance table

Source	DF	Sum of Squares	Mean Square
Total	5	426.83	
z_1^2	1	104.17	
z_2^2	1	56.25	
z_3^2	1	2.08	
z_4^2	1	30.25	
z_5^2	1	234.08	

A textbook would give the following analysis of variance table for our randomized block experiment

Source	DF	Sum of Squares	Mean Square
Total	5	426.83	------
Blocks (classes)	1	104.17	104.17
Treatment (books)	2	58.33	29.17
Error	2	264.33	132.17

The mean square column may be looked at as an estimate of variance; here books appear to have a small variance in relation to blocks or error. What we have labeled error is sometimes called residual or block x treatment interaction. Quite frequently we assume (sometimes without justification) that there is no interaction and call the interaction term error (this is clearly subjective). It is noteworthy that we have said nothing about a model or probability, yet we have been able to partition the data so that it can be analyzed.

The usual probability model for our experiment would be

$$y_{ij} = \mu + b_i + t_j + e_{ij}, \quad i = 1, 2 \text{ (classes)}$$
$$j = 1, 2, 3 \text{ (books)}$$

where

y_{ij} = observed score on i^{th} class, j^{th} book
μ = overall effect
b_i = block or class effect
t_j = treatment or book effect
e_{ij} = random error of the i,j^{th} observation.

If we also make the usual assumption that e_{ij} are normal variates with zero mean and constant variance σ^2 which are independent for all i and j, we may construct the usual F tests. The F test for treatments (books) is

$$F = \frac{\text{Book Mean Square}}{\text{Error Mean Square}}$$

$$= \frac{29.17}{132.17}$$

$$= 0.22.$$

There are two distinct practices as to the interpretation of this F value.

1. The Neyman-Pearson decision theory approach is to reject the null hypothesis of no difference in treatments if the observed F is larger than the tabulated F with a prespecified level of significance. For our example above we would accept the null hypothesis of no treatment difference.

2. Find the probability of a value which is larger than the observed F, assuming the null hypothesis to be true. This is the observed significance level, often called the p-level. For our example above the observed level of significance is roughly 0.8. What does this number mean? Fisher apparently took the point of view that if it is large, then perhaps the null hypothesis is true.

In summary, the analysis of variance may be thought of as a partitioning of the total sum of squares in a meaningful way. To this end, we must search for recognizable sources of variation and think hard about the error and the randomization scheme.

Lecture 13: Some Classes of Problems in Statistical Methods

Introduction

For the past two weeks you have been considering the applications of statistics, more commonly called statistical methods. I wish to consider, during this lecture, the question of what statistical methods are all about. It should be quite clear that the central activity of statistical methods is data analysis. It is undoubtedly equally clear that the underlying theory rests heavily upon the theory of probability; however, I would not wish to go so far as to say that all statistical analysis must be based upon probability models. In order to organize our thinking about the nature of statistical methods, I will proceed as follows. First, I would like to describe what I consider to be several real problems arising in research, development, and business. Secondly, I wish to label statistical problems in statistical methods as falling under one of three very broad classes. Finally, I wish to make some comments about my viewpoint on where I think statistical methods are heading.

Some "Real Life" Problems

Problem Number 1

The manager of a product line department in an electronics company has instructed his inspection department to install procedures to guarantee that the outgoing product will be of high quality. The chief inspector has put into effect an acceptance sampling plan which calls for taking a sample from each lot produced. The sample is tested and the lot is either rejected or accepted, depending upon the test results from the sample. The problem is how to devise a satisfactory sampling plan and decision rule for accepting or rejecting the lot.

Problem Number 2

A battery manufacturer has signed a contract to supply the fuel cells for a generation of space vehicles, conditional upon meeting the reliability requirements. The contract calls for the manufacturer to establish at a 95%

confidence level that there is a .9999 probability that any given fuel cell will meet the engineering specifications called for by the designers of the space vehicles. The problem confronting the battery manufacturer is to devise a sampling scheme and statistical analysis which will convince his customers that he has met the reliability requirements.

Problem Number 3

A psychology researcher wishes to compare the relative effects of four different types of psychological reinforcement upon the performance of certain manual tasks. For his work he has available volunteer school children from three different elementary schools as well as four graduate students to assist him in administering the reinforcement. His problem is to design the study to compare the different types of reinforcement, adjusting as far as possible for the effects of different schools, different classes, and different workers. He also has the problem of analyzing the resulting data in a meaningful fashion.

Problem Number 4

An English professor has become interested in the possible association between English ACT scores of entering college freshmen and their resulting course score in their first course in English composition. He has obtained the English ACT scores and the final course scores for a freshman class of 723 students. He now wishes to characterize in some way the association between the two scores.

Problem Number 5

A research laboratory of a rubber company has been conducting research upon the tear strength of a particular rubber product. They have conducted an experiment in which they have controlled the percentages of components A and B and have then measured the tear strength of the resulting rubber material. Using the data obtained from the experiment, they wish to devise a method of predicting the tear strength, given the percentages of components A and B.

Problem Number 6

A farm economist is studying the trends and fluctuations in pork prices in the State of Oklahoma. He has a record of these prices for the past thirty years. He wishes to perform some sort of analysis which will provide him not only with some tenable theories about what happened in the past, but which will also enable him to predict pork prices for the immediate future.

Problem Number 7

The city planner of a small town is making an effort to foresee the future transportation needs of the city in which he is employed. As a first step, he wishes to determine the load on the main traffic arteries of the city. He has authorized a survey to estimate the average vehicles per day on each of the streets of interest. The natural desire is to obtain a result that is as close as possible to the true figure.

Problem Number 8

An electronics manufacturer measures and records 23 different variables or parameters on each of the electronics components produced. Recognizing the interdependence of many of these measurements, a process engineer suggests that it should be possible to combine many of the variables and parameters and to reduce the number of variables measured to a much smaller set, without the loss of any information. He has measured all 23 variables on a total of 115 components and wishes to analyze the data to see whether it is possible to reduce the number of variables.

Problem Number 9

A researcher in zoology has recorded 15 physical measurements on 100 rats each from two recognized species. Upon the basis of any one measurement, it is impossible to distinguish between the rats of the two species. He is interested in whether it is possible to devise some sort of index which takes into account all the measurements simultaneously and allows one to distinguish between the two species, purely upon the basis of the physical measurements.

Not all problems encountered in the application of statistical methods are represented by this list; however, we believe that the list does span a major portion of the fields of application. In some senses the problems seem to be quite different and the appropriate methods seem to be unrelated. In the next section, however, we will suggest that all of these problems fit within one or more of three general classes of problems.

Some Important Classes of Problems in Statistical Methods

If one were to classify the problems discussed in the previous section by titles commonly encountered in the tables of contents or indexes of most books in statistical methods, we suggest that the nine problems would probably be classified by the following nine titles:

1. Acceptance Sampling
2. Reliability Analysis
3. Design of Experiments
4. Correlation Analysis
5. Multiple Regression
6. Time Series Analysis
7. Survey Sampling
8. Factor Analysis
9. Discriminant Analysis

These labels are useful and do serve to delineate certain areas of work in statistics; however, as can be readily perceived, this is only a partial list. A complete list of the problem areas in statistical methods is too long to permit comprehension of what statistical methods is all about. We would like to suggest that statistical methods can be classified as belonging to one or more of the following three classes of problems:

1. Fitting of models to data
2. Goodness of fit of completely specified models
3. Goodness of fit of partially specified models

We believe that these three subjects provide a useful framework by which one can think intelligently about problems in statistics and solutions to such problems. Clearly, the central feature which characterizes modern statistics is the use of a mathematical or probabilistic model as the basis for the analysis of data. It seems almost axiomatic, then, that the central problems are the fitting of such models to the data obtained and the judgment of how well the models fit the data. In judging the goodness of fit of models to data, it has seemed useful to us to distinguish between completely specified models and partially specified models. In most serious statistical applications we use only partially specified models, although completely specified models are usually used in introducing basic statistical techniques.

Let us now consider the problems presented in the previous section and consider how they may be viewed as falling under one or more of these general classes of problems:

Problem Number 1

The model likely to be assumed for a sample drawn from a production process is that of independent Bernoulli trials with constant probability p. The occurrence of a defective item corresponds to a Bernoulli trial outcome of zero. The problem of estimating the lot fraction defective is then viewed as a problem of estimating the p of the Bernoulli trials, given the acceptance sampling results. The model which we are using is only partially specified, since the p is regarded as unknown. The problem of deciding whether to accept or reject the lot can be viewed as one of comparing the goodness of fit of the sample data with a class of models with p large as opposed to a class of models with p small. If the sample data are better fitted by a partially specified model with unknown p, but smaller than some given value, then we will be led to accept the lot. If, on the other hand, the sample data are better fitted by a partially specified model with large p, then we will be led to reject the lot.

Problem Number 2

A model for the reliability problem might be that the items sampled by the manufacturer constitute a random sample from some mathematical distribution, the degree of specification depending upon whether or not the form of the distribution is assumed. Generally, even if the form of the distribution is assumed, the model is still only partially specified because the parameter values of the distribution will not be assumed to be known. In order for the reliability analysis to have real weight, and to be truly convincing, there must be some indication that the model is an appropriate one. We are therefore led again to the problem of fitting the unknown parameters of the model to the sample data. The question of whether or not the process reliability meets the contractual requirements can be viewed as a matter of the goodness of fit of the data to a partially specified model. We are interested in what fraction of the population exceeds a given engineering requirement. We therefore set out to determine the goodness of fit of the data by a mathematical distribution, only partially specified, in which the area above the engineering specification exceeds the required reliability.

Problem Number 3

In discussing the problem of comparing four different reinforcement conditions, I am going to assume that all of you are familiar with the analysis of variance and are familiar with the use of an F test, given by the treatment mean square over the error mean square, to test the hypothesis that the treatment means are all equal. This sort of test is usually justified by the use of a model which involves an overall effect, a treatment effect, a block effect plus other identifiable effects, plus an error term which is normally and independently distributed with mean zero and constant variance σ^2. I would like to suggest that the F test for no treatment differences can be viewed simply as judging how well the data are fitted by a model in which there are no treatment terms. The model is a partially specified model, of course, because the values for the overall effect, the block effects, and the error variance are not specified.

Problem Number 4

In looking at the ACT scores and the final course scores, the model for the data might be that the pairs of scores constitute a random sample from a bivariate normal distribution with unspecified means, variances, and correlation coefficient. The model is then fitted to the data by the process of what is commonly called estimation of parameters, and judgments about the degree of association between the two measured scores may be viewed as comparing the goodness of fit of the data to a bivariate normal with zero correlation with a bivariate normal with a nonzero correlation.

Problem Number 5

In trying to get an equation to predict the tear strength from input variables, it is very clear that this problem fits in the first class of problems described, namely, that of fitting models to data. It should also be clear that one of the major problems in regression analysis is the forming of opinions about how well the equations fit the data. Since the terms in the regression equations, as well as the value of the error of variance, are generally unspecified, the model is again an example of a partially specified model.

Problem Number 6

The degree of specification for time series analysis varies widely from a model in which we assume only something like variance stationarity or covariance stationarity to a model in which the basic structure (apart from unspecified constants) for the data-generating mechanism is assumed. In the loosely specified case the fitting of the model to the data usually amounts only to the calculation of some simple statistic, such as lag correlation coefficients, or a spectrogram. In the more completely specified models the fitting of the model to the data will involve estimation of the unspecified parameters in the data-generating model. In all cases we must form some opinion about the goodness of fit of the data to the model assumed.

Problem Number 7

In the case of a sample from an existing population, the model is very loosely specified. It usually consists only of the assumption that the individuals in the sample constitute a random sample of a fixed finite population. In this case fitting of the model to the data usually consists only of the calculation of simple things such as the mean or the variance. The goodness of fit may involve problems like judging whether or not the population total exceeds a given value. For example, in the traffic survey referred to, one might be interested in whether or not the data is well-fitted by a model with a traffic load greater than a specified value.

Problem Number 8

If a factor analysis is done without consideration of some underlying model, it remains purely and simply descriptive statistics and only falls marginally within the limits of this lecture. If, however, there is an underlying probability model, then the problems of factor analysis are again those of fitting the sample data to the model and judging the goodness of fit to the model.

Problem Number 9

The model for discriminant analysis may be that the sample data have come from two different mathematical distributions with unspecified parameters. Upon the basis of the data, we wish to fit these distributions to the data, and the resulting problem of discrimination is then a matter of which distribution best fits which observations.

Perspective

To attempt to describe the whole of statistical methods as falling within three classes of problems, is in one sense, foolhardy. Obviously, the description of all statistical activity is an immense task and is, in itself, a statistical problem. However, we believe that the framework we have presented does provide a good perspective for comprehending the whole of the application of statistical methods. We believe that it is consistent with the ideas of the scientific method. We believe that it is useful in helping people to conceive of statistical methods not as all-powerful, but as slightly ad hoc. When statistical methods are viewed in this fashion, rather than as all-purpose remedies prescribed by your local statistician, it becomes apparent that it is very difficult to do a good and effective job in the application of statistical methods.

What about people with insufficient training who have this broad perspective? What about people who wish to use some prescribed statistical methods for the set of data which they have in hand? Should they be encouraged to apply statistical methods in such a cookbook fashion? There are those who would say no. However, it is my own opinion that the routine application of statistical methods, many times by people with only marginal understanding of what they are about, has proven to be useful for scientific investigation. My answer is, therefore, that they should be encouraged to apply statistical methods.

However, as research workers become better trained, they will no longer be satisfied with routine approaches to the analysis of data. It is my personal belief that they will be better satisfied in looking at statistical methods as I have described them here today, even though the resulting analysis may seem to have something more of an ad hoc nature.

I believe that in future years we will see far more statistical analysis than we have witnessed to date. I believe that the users of statistics will become more and more sophisticated and that the general philosophy and attitude toward the application of statistical methods will necessarily change and become more sophisticated.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 14: Regression Analysis

Today I would like to consider regression analysis with respect to:

1. Linear Models
2. The Method of Least Squares
3. Normal Equations
4. Inference about the Parameters of the Linear Model
5. Goodness of Fit of the Model

Linear Models

A typical observation is usually expressed as

$$y_i = f(\text{known } x\text{'s, unknown } \beta\text{'s}) + e_i,$$

where e_i is random error which generally has various assumptions attached. If the function f is a first-degree polynomial in the β's, we call this a linear model. Examples of such linear models are

$$(1) \qquad y_i = \beta_o + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, 2, \ldots, n,$$

or

$$(2) \qquad y_i = \beta_o + \beta_1 \sin x_i + \beta_2 \cos x_i + e_i, \quad i = 1, 2, \ldots, n.$$

Notice that these models are linear in the β's, not in the x's. The set of n observations would then be written out for case (1):

$$y_1 = \beta_o + \beta_1 x_1 + \beta_2 x_1^2 + e_1$$

$$y_2 = \beta_o + \beta_1 x_2 + \beta_2 x_2^2 + e_1$$

$$\vdots$$

$$y_n = \beta_o + \beta_1 x_n + \beta_2 x_n^2 + e_n$$

and for case (2):

$$y_1 = \beta_o + \beta_1 \sin x_1 + \beta_2 \cos x_1 + e_1$$

$$y_2 = \beta_o + \beta_1 \sin x_2 + \beta_2 \cos x_2 + e_2$$

$$\vdots$$

$$y_n = \beta_o + \beta_1 \sin x_n + \beta_2 \cos x_n + e_n$$

Using matrix notation, we have for case (1):

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

and for case (2):

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \sin x_1 & \cos x_1 \\ 1 & \sin x_2 & \cos x_2 \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ 1 & \sin x_n & \cos x_n \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

The general form then for the linear model is the matrix model of the form

$$Y = X\beta + e,$$

where Y is the $n \times 1$ vector of observations, X is the $n \times p$ coefficient matrix of the β's, β is the $p \times 1$ vector of the unknown β's, and e is the $n \times 1$ vector of random errors. From this model of our observations we now want to estimate the β's. The usual technique used is the method of least squares.

The Method of Least Squares

The method of least squares is usually used for two reasons: it is very tractable and it allows the use of a probability model. The least squares solution is found by minimizing with respect to the β's the sum of squares of deviations from the fitted model. We have the very important Gauss-Markoff Theorem:

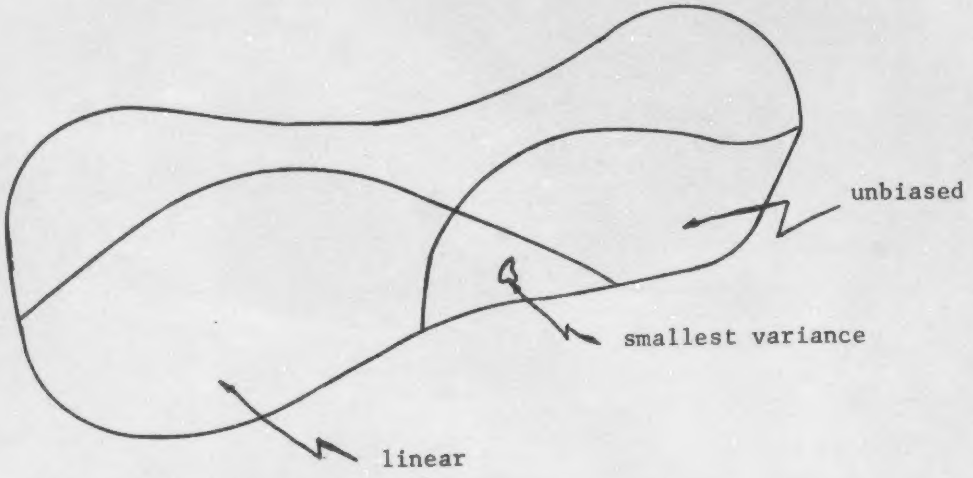
> If the e_i's are independently distributed with zero mean and variance σ^2 and if X has rank p ($p \leqq n$), then the least squares solution $\hat{\beta} = (X^TX)^{-1}X^TY$ is optimal in the sense that each component, say $\hat{\beta}_i$, is a minimum variance linear unbiased estimate of β_i (the independence assumption may be dropped under special conditions).

Let us look at these properties of our estimates $\hat{\beta}_i$.

Linear: We will consider only linear combinations of the observations for estimators.

Unbiased: $\hat{\beta}_i$ is said to be an unbiased estimate of β_i if $E(\hat{\beta}_i) = \beta_i$.

Minimum variance: We will select as our estimate from those which are linear and unbiased that with the smallest variance. Graphically we have the following situation.

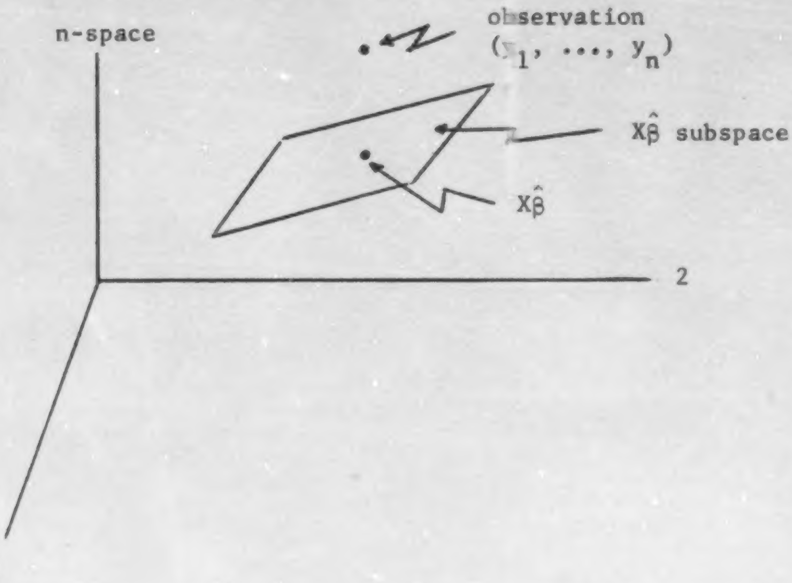


ESTIMATES OF β_i

The Gauss-Markoff Theorem is very important classically, but it is not as strong as it appears at first. The individual estimates are optimal in a sense, but individual optimality does not seem compelling when we consider the joint estimation of the β's. Let's return to the model

$$Y = X\beta + e.$$

We have, in words, that the random variable Y is equal to its mean $X\beta$ plus deviations e from the mean. The vector Y lies in an n-dimensional space. The vector $X\beta$ lies in a p-dimensional subspace of this n-space.



We want to pick $\hat{\beta}$ in the $X\hat{\beta}$ subspace which is, in some sense, "closest" to the observed vector Y. We select $X\beta$ so that $Y - X\hat{\beta}$ is orthogonal to the $X\beta$ plane. This is equivalent to having $Y - X\hat{\beta}$ orthogonal (or normal) to $X\hat{\beta}$. This gives rise to the equation

$$(X\hat{\beta})^T(Y - X\hat{\beta}) = 0$$

or

$$\hat{\beta}^T(X^TY - X^TX\hat{\beta}) = 0,$$

which requires

$$X^TX\hat{\beta} = X^TY.$$

This is the so-called normal equation, where normal implies orthogonal.

It is interesting to note that electrical engineers also work with $X^TX\hat{\beta} = X^TY$ but do not use the terminology of normal equations. They observe a set of vectors $X_1, X_2, \ldots, X_n$ in time and want to predict X_{n+1}. Let $[X_1, X_2, \ldots, X_n] = X$. Then the linear combination of the vectors

$$X_{n+1} = X\hat{\beta}$$

is called a linear filter. They then use the orthogonality principle, i.e., the error of prediction should be orthogonal to the data. The error is $X_{n+1} - X\hat{\beta}$; this is to be orthogonal to X, so

$$X^TX_{n+1} - X^TX\hat{\beta} = 0$$

or

$$X^TX\hat{\beta} = X^TX_{n+1}.$$

This is what we called the normal equations. It should be noted that X_{n+1} is not observed as was our Y.

Statistical Inference about the Parameters (β's)

Let us consider a test of hypothesis

$$H_0: \beta_1 = \beta_3 = \beta_{11} = 0$$

against

$$H_A: H_0 \text{ not true } (\beta_1^2 + \beta_3^2 + \beta_{11}^2 \neq 0, \text{ for example}).$$

How does one test this? Well, we want to see whether β_1, β_3, and β_{11} have any effect on the model. A reasonable thing to do is to put β_1, β_3, and β_{11} in the model and then leave them out and see whether there is any difference. How do we determine whether there is any difference? Let us fit the full model (include β_1, β_3, and β_{11}) and look at the residual sum of squares:

$$\begin{aligned}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) &= Y^TY - \hat{\beta}^TX^TY - Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} \\ &= Y^TY - \hat{\beta}^TX^TY - \hat{\beta}^T(X^TY - X^TX\hat{\beta}) \\ &= Y^TY - \hat{\beta}^TX^TY\end{aligned}$$

This residual sum of squares for the full model is then:

$$(\text{Residual SS})_{\text{Full}} = \sum y_i^2 - \hat{\beta}^T \text{ times the right hand side of the normal equations}$$

To fit the reduced model we just omit the parameters β_1, β_3, and β_{11} (hypothesize that they are zero) and repeat the above procedure. The residual sum of squares for the reduced model is denoted by $(\text{Residual SS})_{\text{Reduced}}$. Note that this will be larger than $(\text{Residual SS})_{\text{Full}}$ because one cannot fit as well with fewer parameters. To compare the differences, it is reasonable to consider

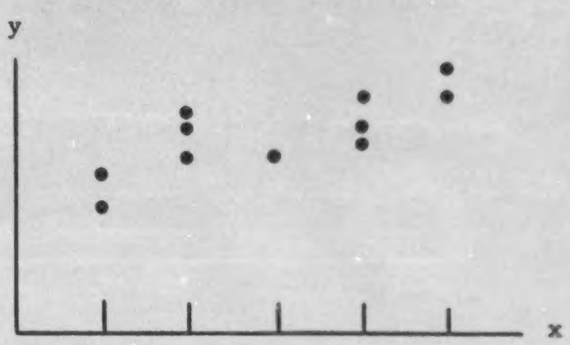
$$\frac{(\text{Residual SS})_{\text{Reduced}} - (\text{Residual SS})_{\text{Full}}}{(\text{Residual SS})_{\text{Full}}} = F^*.$$

Now if the e's are independent normal random variables with zero mean and variance σ^2, then $[(n - p)/3]F^*$ has the F distribution under the null hypothesis. F^* has some appeal as a test without this assumption; with the assumption we may make a probabilistic test. Of course, there are better computational schemes, and much more could be said about this test, but let's proceed to the last topic.

Goodness of Fit of the Model

Checking the goodness of fit of the model is certainly an unresolved issue. In 2-space we can plot the data and certainly get some insight as to whether the model should be linear, quadratic, etc. This is not easily done in higher dimensions.

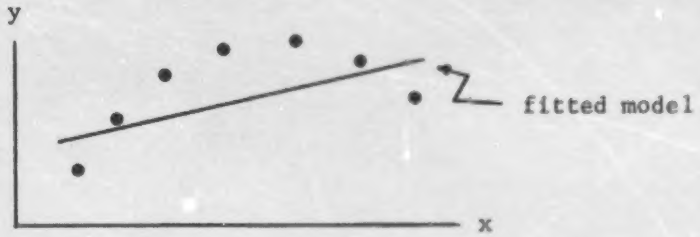
If we have repeated y's for the same x value (or several x values as depicted), we can perform an exact test.



Here we can partition the residual sum of squares into two parts:

1. Within group sum of squares--a measure of variability or estimate of σ^2
2. Remainder sum of squares--a measure of variability plus the lack of fit

As an example, let's consider the extreme situation depicted below.



The residual sum of squares of the fitted model measures only the lack of fit as there is no error present. If error had been present when we fit the line then the residual sum of squares would have contained both lack of fit and error. We may now, with our normal assumption, set up an F test for lack of fit.

If we do not have repeated y's, there is no measure of lack of fit. What we can do here is to look at the residuals, the elements of $Y - X\hat{\beta}$. One way to do this is to plot the residuals versus fitted values $\hat{y}$. The plot should appear fairly horizontal, as the error should not depend on the magnitude of the response. We see this from the assumptions of zero mean and constant variance σ^2 for our error term. If the plot trends (upward, say), this indicates that perhaps we should transform our data, as our assumptions are violated. This leads us to the subject of probability plotting. Here we could plot the residuals versus their cumulative distribution function; however, use of special graph paper changes the scale so that such a plot appears linear. Such plots are very useful in checking the normality assumption.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 15: A Theory of Statistical Inference

Introduction

Today we will continue our discussion of the application of statistics. Previously I discussed a mode of thinking for the whole of statistical methods. The concept of statistical inference was not discussed explicitly, but certainly it lies at the foundation of the application of statistics.

I believe that applied statistics is presently in a very difficult, although interesting, time. When I was studying statistics as a graduate student, I believed that there was a universal philosophy of statistics generally accepted by most members of the profession. Under this universal viewpoint, statistical principles seemed good or bad, true or false. Today things seem to be more confused and more muddled. I think this is evidenced by the fact that our colleagues in other fields are also confused and puzzled. I hear such comments as "Can I say that I am 95% confident that the mean is in a given interval after the sample is drawn?" "What is this Bayesian statistics?" "Is it true that the Bayesian estimates of regression coefficients are more powerful than other methods?"

I believe that these and other similar questions are motivated by divergent views about a subject at the very foundation of all statistical methods, namely, statistical inference itself. For about thirty years, the Neyman-Pearson school of thought on statistical inference has dominated statistics. The Neyman-Pearson theory is still the theory most commonly presented in books on mathematical statistics. However, in more recent times we have heard the idea that all statistics should be viewed as decision-making. Others propound the use of Bayes' Theorem as the panacea for all problems. Regardless of how one looks at all of these ideas, it is very clear that there is not a single school of thought on the subject of statistical inference. In a paper published in the Journal of the Royal Statistical Society in 1968, Professor M. G. Kendall expressed the following sentiment:

> "What I do mean is that neither the Pearsonian nor the Fisherian movement is any longer pushing out the frontiers of our subject. We are at the end of one era. Whether we stand at the beginning of another, in the sense that developments in the next two or three decades can be

grouped around a single strand of thought, I rather doubt. We are not heading in any one particular direction and, consequently, it is much more difficult to predict the future lines of development than it was twenty-five years ago."

Because of its fundamental role at the foundation of applications, because of the many different philosophies of statistical inference, and because of the present uncertainty over the direction in which we are moving, I felt that it would be important to discuss with you today some theories of statistical inference, including my own.

What is Statistical Inference?

Wanting to give something other than the stock answer that statistical inference is estimation, confidence intervals, and hypothesis testing, I ask you to consider the name "statistical inference." One dictionary definition of inference is "The act of inferring." The same dictionary defines infer, "To deduce or accept on the basis of evidence." Statistics is defined as "Quantitative data." It seems, then, that we would be led to define statistical inference as the process by which one forms opinions from the facts available. Although such a description of statistical inference seems to follow naturally enough from the definition of the component terms, the definition does seem a little ambitious. It would appear that statistics embraces all empirical science in the sense that all empirical science is concerned with trying to find out things from the facts available. On the other hand, one might be led to ask, "How else would one find out things without drawing conclusions from the facts available?" In this sense, the definition seems unnecessarily limited.

There is one aspect of statistical inference which I wish to emphasize in passing. Focusing upon the conclusion-making aspects of statistical inference, some statisticians have come to regard all statistical inference as a problem of decision-making. While it may possibly be true that every activity of the human mind can be viewed as a process of decision-making, I feel that it is not fruitful to do so. I prefer to make a distinction between a process of decision-making and a process of forming opinions. I realize that this distinction is not sharp, but I feel that it is worthwhile because analysis appropriate for making decisions may be totally inappropriate for forming opinions about the data. For example, a toss of a coin may be a meaningful part of a decision-making process, but it is totally uninformative when it comes to forming opinions about the data.

Statistical Decision Theory

I shall not attempt to give a careful formulation, but shall say that in general we are concerned with a choice of a decision rule which maps our data into a decision or action. The choice of a rule depends upon our consideration of the consequences of the action and the state of nature. This is generally formalized through the consideration of a loss function which is a function of the decision and the state of nature. Since the decision depends upon the data, the loss function is a function of the data and the state of nature. Although some simple problems and decision-making can be considered by looking at the loss function, we generally pass from the loss function to what is called the risk function. The risk function is taken to be the average over all possible data sets of the loss function. Therefore, the risk function is a function of the possible decision rule and the state of nature. Once we are led to the risk function, we must then choose a decision rule which optimizes, in some way, the value of the risk function. A conservative rule which has been put forward is the minmax rule. The minmax rule minimizes, over all possible decision functions, the maximum risk taken over all possible states of nature. A rule of considerably more appeal, and also more controversy, is the Bayes rule. The Bayes rule considers a weighted average risk, the weights being assigned to the different possible states of nature, and the Bayes decision rule is the decision rule which minimizes the weighted average risk. The controversy arises over the choice of weights and the interpretation of weights. The weights are commonly called the prior probabilities for the state of nature. Given that the prior distribution is commonly accepted, there is no controversy. However, the existence or nonexistence of a prior distribution on the states of nature depends very much upon how one views and interprets probability.

It seems to me that many of the problems in statistics are clearly decision-making problems and, if at all possible, the techniques of statistical decision theory should be applied to their solution. For example, problems in reliability, quality control, and acceptance sampling are very often decision theory problems. Further, in some such problems, all the basic ingredients necessary for the routine application of existing statistical decision theory are available. In terms of the classes of statistical problems which we discussed previously, it seems to me that the first class, which is the fitting of models to data, is a class of decision-making problems. The problem of fitting a model to data is a problem of deciding what values should be used for the parameters in the model.

Statistical Tests

In this section I wish to describe primarily tests of significance. In so doing, I wish to distinguish between tests of significance and tests of hypothesis. R. A. Fisher certainly made the distinction; however, I cannot claim with certainty that my viewpoint is that of Fisher. The earliest example of a statistical test of which I am aware was given by Arbuthnot in 1710. However, the real development of statistical tests began with Karl Pearson, who is known primarily for his development of a chi-square goodness of fit test. Next came W. S. Gossett, who published under the name of Student in 1908. Following the work of Student, R. A. Fisher formulated many tests of significance. In the 1930's Neyman and Pearson put forward a serious theoretical development of a theory of tests. I believe that their initial attempt was more in the spirit in which Fisher conceived of tests of significance. However, as their work proceeded, a statistical test was formalized into a rigid accept-reject rule.

Suppose that we have a set of data. On the basis of prior knowledge, one surmises that a certain class of models can be used to represent the data. We ask whether the data can provide a measure of evidence in support of the given class of models. From this viewpoint, then, a test of significance maps the data into an observed significance level. The observed significance level is a measure of agreement between the data and the hypothesized model. The larger the significance level, the greater is the support for the model given by the data. The smaller the significance level, the less the support.

It may well be that we do not entertain the idea that the hypothesized class of models is the true one. As Fisher stated in 1956, "In general, tests of significance are based on hypothetical probabilities calculated from the null hypothesis. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypothesis they test." We use a test of significance to form an opinion of whether the data conform to the hypothesized model. We believe that in most applications the hypothesized model is unacceptable as a true model; rather, it is set up as a "straw man" to determine whether the data suggest that it is false. For example, a common experiment consists of giving a treatment to one group of subjects and another treatment to another group. The hypothesis that the treatments have equal effects is then tested. In most cases the experimenter knows that this hypothesis is not true. Yet, he wishes to determine whether the data provide some evidence.

Hypothesis testing, on the other hand, is concerned with accept-reject rules. Upon the basis of the data, we decide whether to accept or reject a given model.

Viewed in this way, tests of significance are clearly concerned with inductive inference, while tests of hypothesis are concerned with inductive behavior.

Statistical Intervals

Having taught statistics for a number of years and always having had difficulty with the presentation of confidence intervals, I mention this topic with some timidity. Probably you are familiar with the fact that confidence intervals can be obtained from tests of hypothesis by a process which we shall call inversion. For example, the usual 95% confidence interval on the mean of a normal distribution can be obtained by determining the values of μ which would be accepted at the 5% level of significance. Tracing the probability argument in the derivation of the test of hypothesis enables us to make the statement that we are 95% confident that the interval contains the mean. In addition to the considerable difficulty of explaining the meaning of this statement and in cautioning the student to realize that the probability holds before the sample is drawn and not after, I usually have some difficulty in convincing the student that the confidence statement has any great utility. Perhaps the difficulty is that the student expects and wishes the statement to be stronger than it really is.

We have presented tests of significance as being logically different from tests of hypothesis. Of course, the calculations in any given problem usually will be about the same; however, the underlying philosophy will be totally different. What about intervals obtained from inversion of tests of significance? Although we may calculate exactly the same interval, our viewpoint about the interpretation of the interval may be totally different. Instead of saying that 95% of the time the interval calculated will contain the true value

373

of the mean, the statement comes out something like: the values of μ within a given interval are supported by the data as indicated by a certain test of significance with significance levels greater than or equal to .05. Thus, we conceive of the inversion of a test of significance given an interval of values more as a data analysis technique than as a decision-making device.

Bayesian Inference

Any discussion of statistical inference would be incomplete without a brief note on Bayesian inference. Any discussion in depth of this topic must necessarily become involved with one's interpretation of probability. However, we shall not discuss this matter in this lecture. In the simplest form of Bayesian inference, one conceives of the parameters in the model as random variables themselves with a known marginal probability distribution. This known marginal probability distribution is referred to as the prior distribution. Given a set of data, one then uses Bayes' Theorem to calculate the conditional probability distribution of the parameters, given the observed data. This conditional distribution is referred to as the posterior distribution of the parameter values. If one has confidence in the model and in the prior distribution, then the posterior distribution certainly seems to be a meaningful quantity. If, on the other hand, the confidence in the model and prior distribution is somewhat shaky, the posterior probability of a given parameter value may be viewed with somewhat more skepticism.

Conclusion

In this short lecture on theories of statistical inference, we have skimmed very lightly over some very deep, sophisticated, and subtle ideas. In so doing we have undoubtedly done injustice to some of the viewpoints. However, perhaps we will have motivated you to read and study further.

References

1. Fisher, R. A. Statistical Methods and Scientific Inference. London, Oliver and Boyd, 1956.

2. Kendall, M. G. "On the future of statistics--a second look." Journal of the Royal Statistical Society, Series A, 131, 1968, pp. 182-294.

APPLICATIONS OF PROBABILITY AND STATISTICS

Lecture 16: Modeling of Tree Nut Removal

Introduction

Background

Macadamia nut is a delicious and expensive tropical tree nut grown commercially in Hawaii. The harvesting season for macadamia nuts generally exceeds four months. During this period both mature and immature nuts can be found on the trees.

Vibratory detachment of nuts by a tree shaker presents a possible way to mechanize the harvesting operation. Economical feasibility of vibratory nut removal depends on two criteria:

(1) What percentage of the mature nuts is removed from the tree?

(2) What percentage of the immature nuts is removed from the tree?

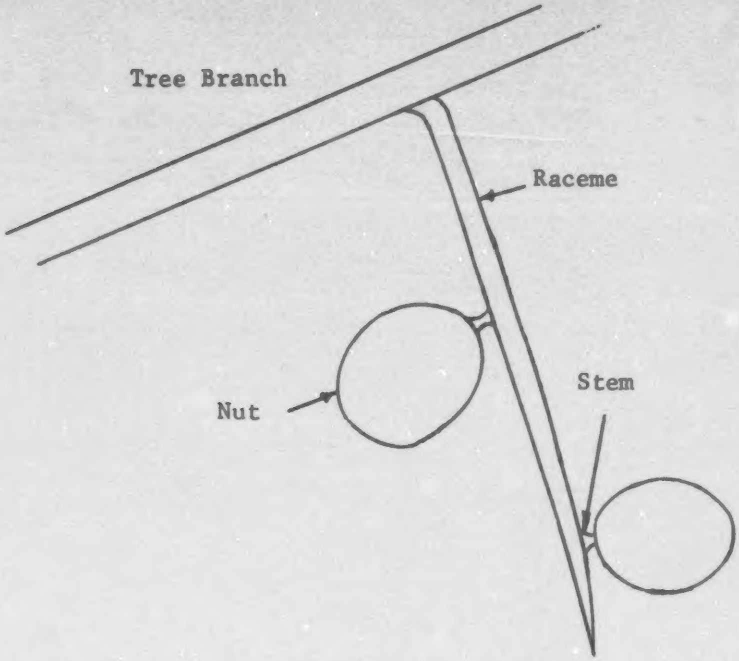
Obviously, at the time of harvest we wish to remove as many mature nuts as possible and remove as few immature nuts as possible.

While a macadamia tree may grow to 50 feet tall, in order to simplify our thinking we may regard it as a collection of nut-raceme systems. Figure 1 shows a typical nut-raceme system.

Raceme length varies from less than an inch to more than 7 inches. Figure 2 shows the distribution of raceme lengths.

There is an indication that the response of a nut-raceme system to an applied frequency depends, among other factors, on the raceme length and nut maturity. Figures 3, 4, and 5 show the differences in mature and immature nut-raceme system responses to an applied frequency of 800 cycles per minute (cpm). There is reason to believe that in order to achieve optimal harvesting (high percentage of mature nuts removed, low percentage of immature nuts removed) a combination of several vibratory frequencies, each applied for different durations, should be employed.

Based on a paper by Tung Liang, D. K. Lewis, Jaw-Kai Wang and G. E. Monroe entitled "Random Function Modeling of Macadamia Nut Removal by Multiple Frequency Vibration." ASAE Paper No. 70-132. Agricultural Engineering Department, University of Hawaii.



MACADAMIA NUT-STEM-RACEME SYSTEM

Figure 1

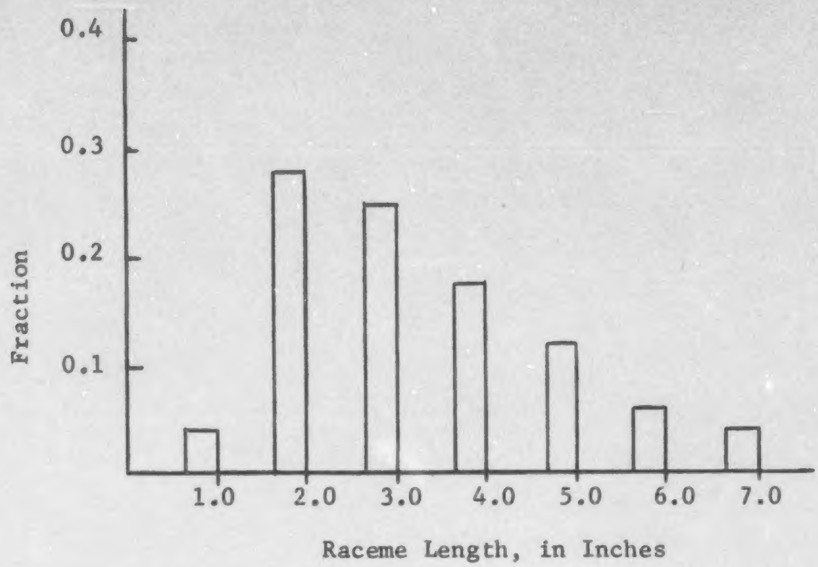
The Problem

The scattering of experimental data has made it necessary to conduct 350 tests using individual nut-raceme systems in order to have sufficient data to construct Figures 3 and 4, which show the response of a random collection of nut-raceme systems to a single applied frequency.

Generally, the frequencies of interest range from 800 cpm to 1,200 cpm. If we allow the total shaking to vary from 0 to 10 seconds at one-second intervals and the frequencies to assume values from 800 to 1,200 cpm at 100 cpm intervals, a sequence of three frequency-duration combinations would be $\{(i_1,k_1), (i_2,k_2), (i_3,k_3)\}$, where i is shaking frequency in cpm, k is shaking duration in seconds, and where $k_1 + k_2 + k_3 = 10$ seconds and i_1, i_2, i_3 may assume any value from the set $\{800 + 100n: n = 0, 1, 2, 3, \text{or } 4\}$.

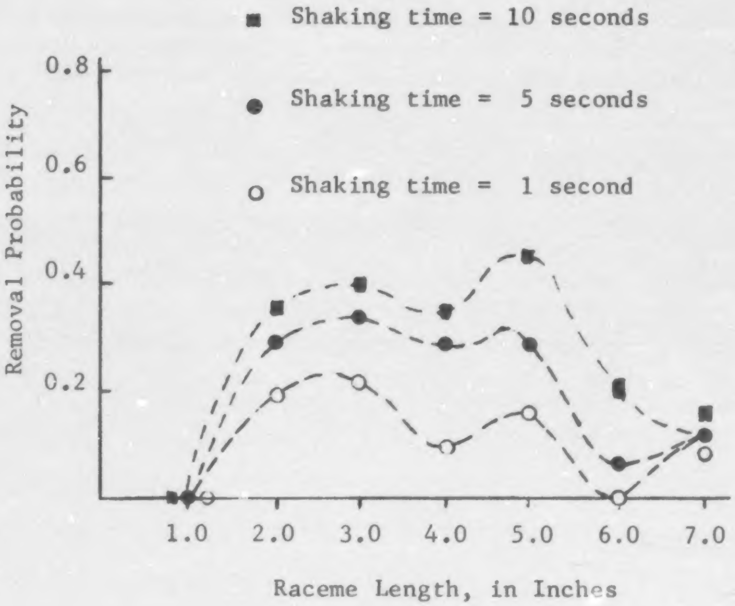
To conduct experiments on each combination with adequate replications would require thousands of tests. It is obvious that some mathematical model is needed to reduce the amount of experimental work.

One possible approach involves the consideration of nut-raceme system responses to single-frequency experimentations, using these to estimate nut removal probabilities of multiple-frequency removal.



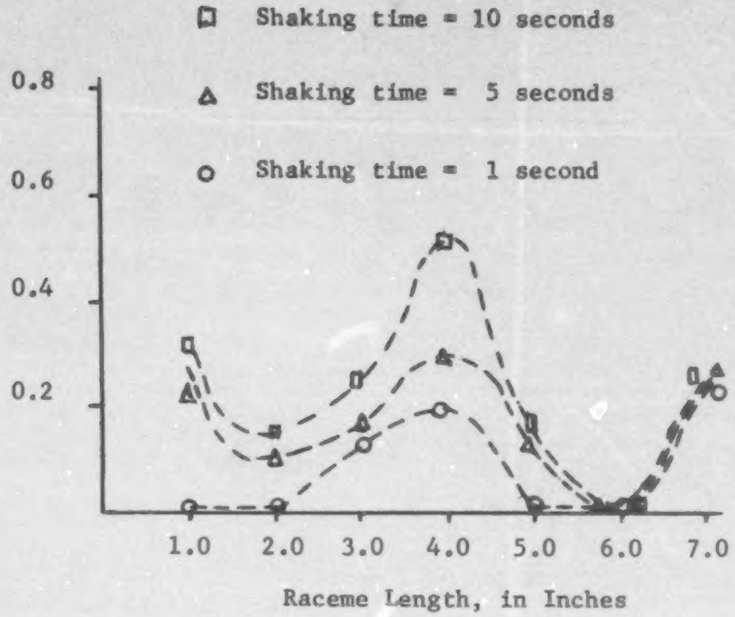
MACADAMIA NUT RACEME LENGTH DISTRIBUTION IN VARIOUS LENGTH INTERVALS

Figure 2



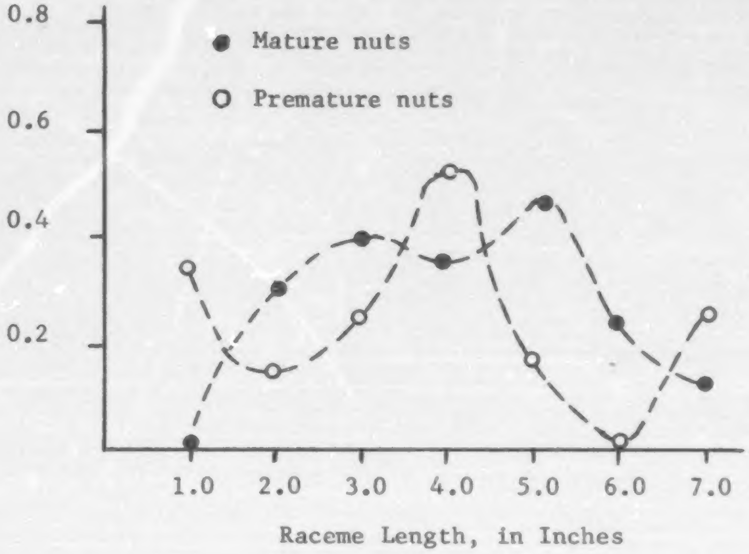
MATURE NUT REMOVAL PROBABILITY VERSUS SHAKING TIME AND RACEME LENGTH AT SHAKING FREQUENCY = 800 CYCLES/MINUTE

Figure 3



PREMATURE NUT REMOVAL PROBABILITY VERSUS SHAKING TIME AND RACEME LENGTH AT SHAKING FREQUENCY = 800 CYCLES/MINUTE

Figure 4



COMPARISON OF THE REMOVAL PROBABILITIES OF PREMATURE AND MATURE NUTS UNDER 800 CYCLES/MINUTE SHAKING FREQUENCY AND 10-SECOND SHAKING INTERVAL

Figure 5

Mathematical Models

Consider the macadamia tree as a random population of nut-raceme systems. Subject this random population first to a shaking frequency of i_1 for a duration k_1, followed by i_2, k_2 and i_3, k_3. The event that a mature nut with raceme j inches long will be removed when subjected to frequencies i_1, i_2, i_3 for k_1, k_2, k_3 intervals, respectively, may be mathematically expressed as

$$\begin{aligned} &\{R(i_1,j,k_1) = 1\} \\ &\cup \left(\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\}\right) \\ &\cup \left(\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 0\} \cap \{R(i_3,j,k_3) = 1\}\right), \quad (1) \end{aligned}$$

where R is a random variable on the set of triples (i,j,k), representing the response of a mature nut with raceme j inches long to frequency i for interval k: $R(i,j,k) = 1$ means that a mature nut with raceme length = j is removed at shaking frequency = i and shaking duration = k; $R(i,j,k) = 0$ means that the nut is not removed.

Now let G be a random variable denoting the response of a premature nut to shaking frequencies and shaking intervals, and let P(A) denote the probability of event A.

The probability P_{rj} that a mature nut with j-inch raceme length will be removed can be expressed as a function of single-frequency responses and some conditional responses. The derivation is given on the next page.

$$
\begin{aligned}
P_{rj} &= P[\{R(i_1,j,k_1) = 1\} \cup \Big(\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\}\Big) \\
&\quad \cup \Big(\{R(i_1,j,k_1 = 0\} \cap \{R(i_2,j,k_2) = 0\} \cap \{R(i_3,j,k_3) = 1\}\Big)] \\
&= P[\{R(i_1,j,k_1) = 1\}] + P[\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\}] \\
&\quad + P[\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 0\} \cap \{R(i_3,j,k_3) = 1\}] \\
&\quad - P[\{R(i_1,j,k_1) = 1\} \cap \{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\}] \\
&\quad - P[\{R(i_1,j,k_1) = 1\} \cap \{R(i_1,j,k_1) = 0\} \\
&\qquad \cap \{R(i_2,j,k_2) = 0\} \cap \{R(i_3,j,k_3) = 1\}] \\
&\quad - P[\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\} \\
&\qquad \cap \{R(i_2,j,k_2) = 0\} \cap \{R(i_3,j,k_3) = 1\}] \\
&= P[\{R(i_1,j,k_1) = 1\}] + P[\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\}] \\
&\quad + P[\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 0\} \cap \{R(i_3,j,k_3) = 1\}] \qquad (2)
\end{aligned}
$$

(The three terms preceded by minus signs are all equal to zero because the intersections are empty events.) Continuing, we have:

$$
\begin{aligned}
P_{rj} &= P[\{R(i_1,j,k_1) = 1\}] \\
&\quad + P[\{R(i_1,j,k_1) = 0\}]\, P[\{R(i_2,j,k_2) = 1\} | \{R(i_1,j,k_1) = 0\}] \\
&\quad + \Big(P[\{R(i_1,j,k_1) = 0\}]\, P[\{R(i_2,j,k_2) = 0\} | \{R(i_1,j,k_1) = 0\}] \\
&\qquad \times P[\{R(i_3,j,k_3) = 1\} | \{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 0\}]\Big) \qquad (3)
\end{aligned}
$$

Similarly, the probability P_{gj} of removing a premature nut with j-inch raceme may be expressed as:

$$
\begin{aligned}
P_{gj} = {} & P[\{G(i_1,j,k_1) = 1\}] \\
& + P[\{G(i_1,j,k_1) = 0\}]\ P[\{G(i_2,j,k_2) = 1\} | \{G(i_1,j,k_1) = 0\}] \\
& + \Big(P[\{G(i_1,j,k_1) = 0\}]\ P[\{G(i_2,j,k_2) = 0\} | \{G(i_1,j,k_1) = 0\}] \\
& \times\ P[\{G(i_3,j,k_3) = 1\} | \{G(i_1,j,k_1) = 0\} \cap \{G(i_2,j,k_2) = 0\}]\Big) \qquad (4)
\end{aligned}
$$

The problem now is to evaluate the conditional probabilities in equations 3 and 4. To do this experimentally would be extremely tedious. Instead, we can estimate the upper and lower bounds of these probabilities.

The upper bound of a conditional probability $P(A|E)$ is given by $P(A)$. Therefore, the upper bound for P_{rj} in equation 3, designated $\overline{P}_{rj}$, is given by equation 5. Similarly, $\overline{P}_{gj}$ is given by equation 6.

$$
\begin{aligned}
\overline{P}_{rj} = {} & P[\{R(i_1,j,k_1) = 1\}] + P[\{R(i_1,j,k_1) = 0\}]\ P[\{R(i_2,j,k_2 = 1\}] \\
& + P[\{R(i_1,j,k_1) = 0\}]\ P[\{R(i_2,j,k_2) = 0\}]\ P[\{R(i_3,j,k_3 = 1\}] \qquad (5)
\end{aligned}
$$

$$
\begin{aligned}
\overline{P}_{gj} = {} & P[\{G(i_1,j,k_1) = 1\}] + P[\{G(i_1,j,k_1) = 0\}]\ P[\{G(i_2,j,k_2) = 1\}] \\
& + P[\{G(i_1,j,k_1) = 0\}]\ P[\{G(i_2,j,k_2) = 0\}]\ P[\{G(i_3,j,k_3) = 1\}] \qquad (6)
\end{aligned}
$$

To obtain the lower bounds of the conditional probabilities in equations 3 and 4, one is first referred to Figure 6.

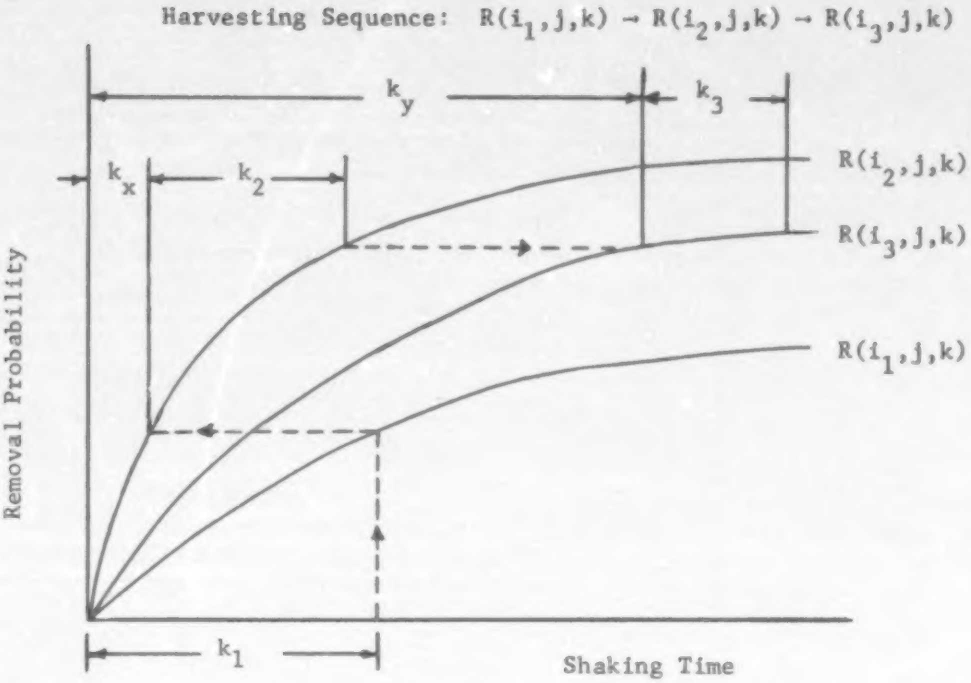
If we assume that the net physical effect on the nut-raceme system with j-inch raceme subjected to (i_1,k_1) is equal to that on the same system subjected to (i_2,k_x), then

$$
\begin{aligned}
& \{R(i_1,j,k_1 = 1\} \cup [\{R(i_1,j,k_1) = 0\} \cap \{R(i_2,j,k_2) = 1\}] \\
& = [\{R(i_2,j,k_x = 1\}] \cup [\{R(i_2,j,k_x) = 0\} \cap \{R(i_2,j,k_2) = 1\}] \\
& = [\{R(i_2,j,k_2 + k_x) = 1\}]. \qquad (7)
\end{aligned}
$$

However, the assumption is not quite accurate for a number of reasons. Equations 8 and 9 will give estimated lower bounds for P_{rj} and P_{gj}. The quantities k_x and k_y in Figure 6 are determined by experimental data.

$$\underline{P}_{rj} = P[R(i_3,j,k_y+k_3) = 1] \tag{8}$$

$$\underline{P}_{gj} = P[G(i_3,j,k_y+k_3) = 1] \tag{9}$$



ESTIMATES OF LOWER BOUNDS OF CONDITIONAL PROBABILITIES FROM SINGLE FREQUENCY RESPONSE CURVE

Figure 6

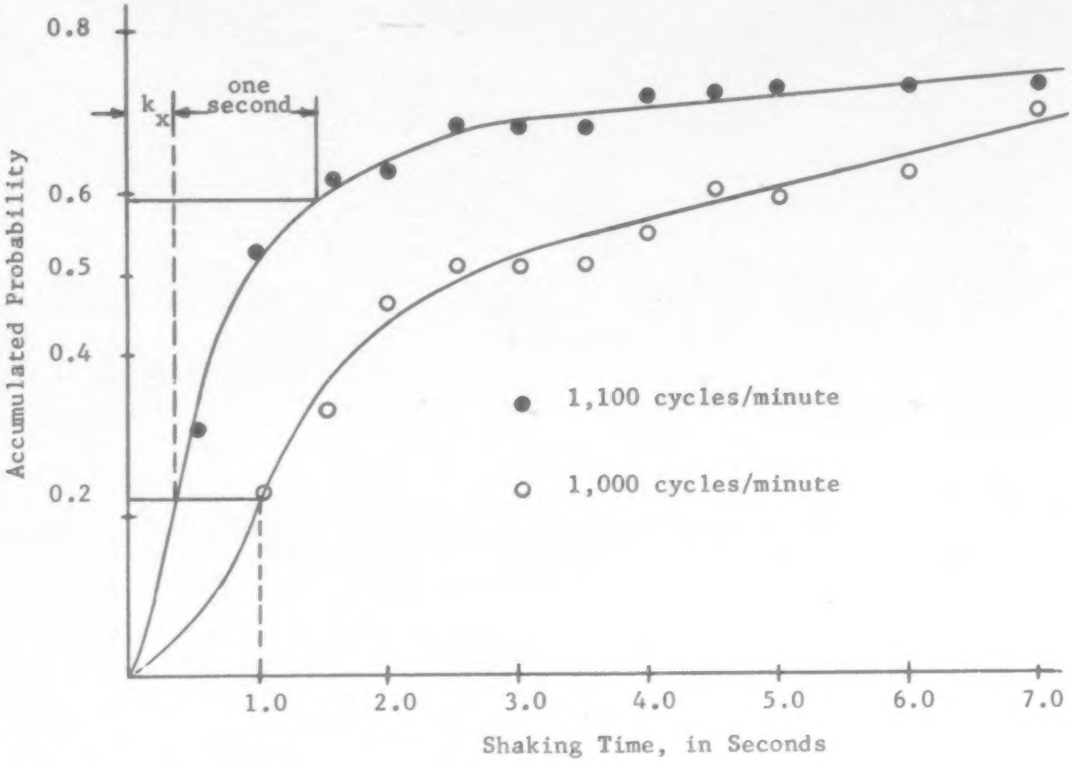
Using equations 5, 6, 8, and 9 we can estimate the effectiveness of multiple-frequency vibration on nut removal using the single-frequency nut removal curve.

Evaluating the Model

Experiments were conducted to evaluate the upper and lower bounds approach. Nut-raceme systems with 2-inch long raceme were shaken at 1000 and 1100 cycles per minute. Samples of 60 nut-raceme systems were used to obtain nut-raceme response to either 1000 cpm or 1100 cpm vibration. Mature nuts were used exclusively. In addition to the nut-raceme systems used for establishing the response to single frequencies, another 250 nut-raceme systems were used to determine the response experimentally to two frequency shakings. They were divided into 25 groups and were subjected to 1000 and 1100 cpm vibration in sequence. The vibration time was one second for each frequency.

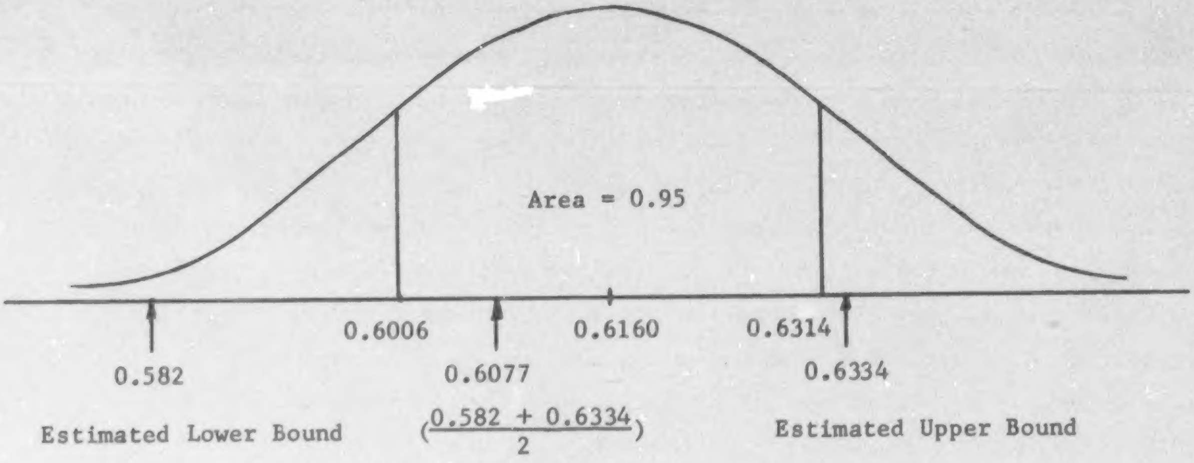
The nut-raceme system responses to 1000 cpm and 1100 cpm, 3/4-inch stroke vibrations are plotted in Figure 7.

The lower bound of the removal probability for the 1000 cpm/one-second and 1100 cpm/one-second treatment is obtained graphically as 0.582. The calculated upper bound is 0.6334. These estimates are compared to the actual experimental data as shown in Figure 8.



MATURE NUT-RACEME SYSTEM TO 1000 AND 1100 CYCLES/MINUTE FREQUENCY SHAKING
(Shaking Stroke = 3/4")

Figure 7



COMPARISON OF MODEL PREDICTED NUT REMOVAL PROBABILITIES AND DIRECT EXPERIMENT RESULT (1000 cpm/one-second followed by 1200 cpm/one-second)

Figure 8

Discussion

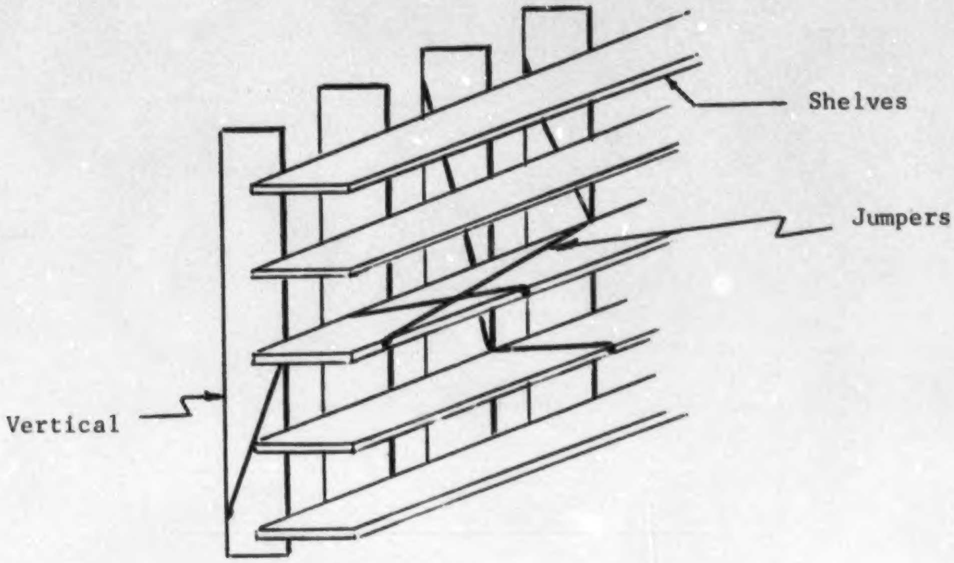
This example shows how mathematical models can be used to reduce tedious experimental work and to obtain reasonable estimations of multiple-frequency nut removal probabilities.

It brings vividly to our minds that exact mathematical solutions are often unnecessary in engineering and that a combination of experimental and mathematical reasoning can frequently serve the interests of engineers best.

Lecture 17: A Telephone Exchange Problem

1.0 Introduction

Every telephone set is connected by a pair of wires to the switching equipment in the telephone exchange office where the communication link between the calling and called party is established. The enormous number of telephone wire pairs enter the exchange in cables. Between the cables and the switching equipment is a large steel structure called the main distributing frame or MDF. The cables are terminated on the MDF on vertical columns of terminals (called verticals) and the equipment is terminated on horizontal rows of terminals called shelves.



MAIN DISTRIBUTING FRAME

Figure 1

Figure 1 illustrates how three different subscribers may be linked by wire pairs (called jumpers) from their termination on a vertical to an equipment termination on a shelf. It is important to note that each terminal on a shelf determines a unique telephone number. Hence, when a subscriber moves within an area served by one exchange his new telephone set may be connected to a different vertical. The procedure is to remove the old jumper and place a new one from the new termination on the vertical to the old termination on the shelf; this assures that the subscriber's number remains unchanged.

The jumper pile-up on the shelves leads to severe congestion and determines the ultimate capacity of an exchange office.

Using assumptions which can be shown to be realistic, we indicate how some simple properties of jumper pile-up can be derived.

2.0 Analysis

2.1 Definitions

A trapezoid S whose two parallel sides are given by

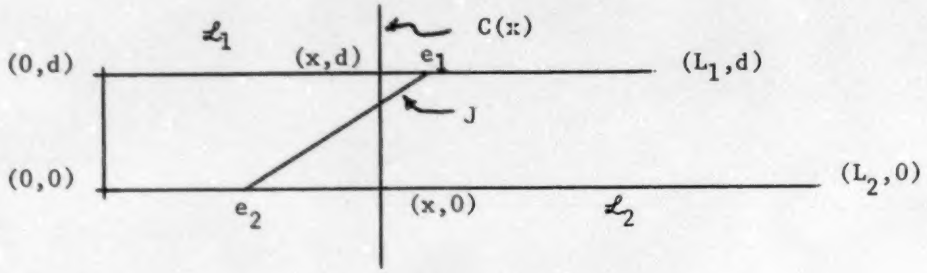
$$\mathcal{L}_1 = \{(x,y): \quad 0 \leqq x \leqq L_1, \ y = d \geqq 0\}$$

$$\mathcal{L}_2 = \{(x,y): \quad 0 \leqq x \leqq L_2, \ y = 0\}$$

will be called a shelf.

Let $e_1 \in \mathcal{L}_1$ and $e_2 \in \mathcal{L}_2$ be two points. The straight line segment $J = (e_1, e_2)$ with e_1 and e_2 as endpoints will be called a jumper.

A straight line C(x) passing through the points $(x,0) \in \mathcal{L}_1$ and $(x,d) \in \mathcal{L}_2$ is defined to be the shelf cross section at x.



Let E_1 and E_2 be two independent uniform random variables with probability density functions

$$f_1(x) = L_1^{-1} \qquad 0 \leqq x \leqq L_1$$

$$= 0 \qquad \text{otherwise}$$

$$f_2(x) = L_2^{-1} \qquad 0 \leqq x \leqq L_2$$

$$= 0 \qquad \text{otherwise}$$

respectively.

2.2 The probability that a jumper traverses C(x), when $L_1 = L_2 = 1$ and $d \geqq 0$

Let

$$P(x) = P\{C(x) \cap J \neq \emptyset\}$$

be the probability that a jumper traverses the shelf cross section at x. Then

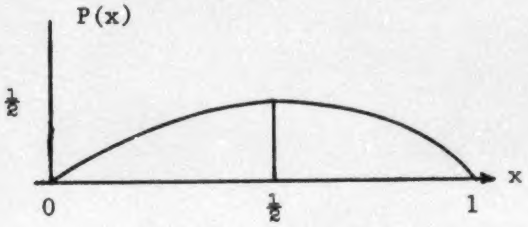
$$P(x) = P\{e_1 \varepsilon ((x,d),(1,d))\} \times P\{e_2 \varepsilon ((0,0),(x,0))\}$$

$$+ P\{e_1 \varepsilon ((0,d),(x,d))\} \times P\{e_2 \varepsilon ((x,0),(1,0))\}$$

$$= (1 - x)x + x(1 - x)$$

$$= 2x(1 - x).$$

It follows that max $P(x) = P(\frac{1}{2}) = \frac{1}{2}$ and $P(0) = P(1) = 0$.



Hence, the maximum jumper pile-up should occur halfway between the frame extremities.

2.3 The probability density function of the jumper length

Let x_i be the distance from the origin to the orthogonal projection of e_i onto the x-axis (i = 1,2). The length ℓ of a jumper is then given by

$$\ell = (|x_1 - x_2|^2 + d^2)^{\frac{1}{2}}.$$

Let L be the random variable of jumper length with $f_L(\ell)$ as probability density function.

We shall derive $f_L(\ell)$ for the case when $d = 0$ and $L_1 = L_2 = 1$.

Then $\ell = |x_1 - x_2|$ and

$$F_L(\ell) = P(L \leq \ell) = P(|E_1 - E_2| \leq \ell).$$

Since each jumper is uniquely determined by the ordered pair of real numbers (x_1, x_2) and conversely each pair of real numbers (x_1, x_2), $0 \leq x_1, x_2 \leq 1$, determines a unique jumper on the shelf, we may view this as a geometric probability problem. Every point in the unit square of Figure 2 represents a jumper, and those in the shaded region whose area is $2\ell - \ell^2$ correspond to jumpers of length less than or equal to ℓ.

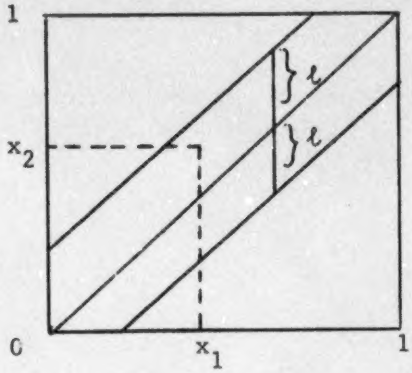


Figure 2

Hence

$$F_L(\ell) = 2\ell - \ell^2 \qquad 0 \leq \ell \leq 1$$

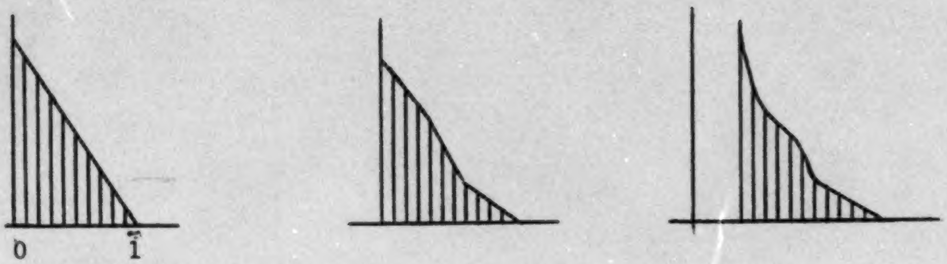
$$f_L(\ell) = 2 - 2\ell \qquad 0 \leq \ell \leq 1$$

$$= 0 \qquad \text{otherwise}$$

and

$$E(L) = 1/3.$$

Figure 3 illustrates the p.d.f. of the random variable L for different assumptions.



$L_1 = L_2 = 1$ $d = 0$

$L_1 \neq L_2$ $d = 0$

$L_1 \neq L_2$ $d > 0$

Figure 3

3.0 Further Problems

3.1 Let $P(x,y;\Delta)$ be the probability that a jumper traverses a small square with sides of length Δ and center $(x,y) \in S$. Then

$$\rho(x,y) = \lim_{\Delta \to 0} \frac{P(x,y;\Delta)}{\Delta}$$

is an approximation to the convex hull formed by the jumpers on a shelf. Here again the representation of jumpers as points in $\mathcal{L}_1 \times \mathcal{L}_2$ is readily applicable. For $L_1 = L_2$ and $d > 0$ the convex hull will have a form as shown in Figure 4.

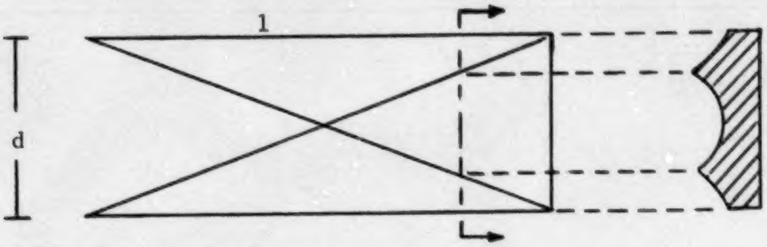


Figure 4

3.2 When $\mathcal{L}_1$ and $\mathcal{L}_2$ are concentric circles with radii R_1 and R_2, respectively, one may proceed along the same lines as before to obtain some useful results. When $R_1 = R_2 = 1$ we get

$$F_L(\ell) = \ell/\pi \qquad 0 \leqq \ell \leqq \pi$$

$$f_L(\ell) = 1/\pi \qquad 0 \leqq \ell \leqq \pi$$

$$= 0 \qquad \text{otherwise}$$

and

$$E(L) = \pi/2.$$

3.3 Given an appropriate convex hull, what are the corresponding p.d.f.'s on L_1 and L_2?